# Typing Performance of Blind Users: An Analysis of Touch Behaviors, Learning Effect, and In-Situ Usage

Hugo Nicolau[1], Kyle Montague[2], Tiago Guerreiro[3], André Rodrigues[3], Vicki L. Hanson[1,2]
[1]Rochester Institute of Technology, [2]University of Dundee,
[3]LaSIGE, Faculdade de Ciências, Universidade de Lisboa
hmnics@rit.edu, kmontague@dundee.ac.uk, {tjguerreiro, afrodrigues}@fc.ul.pt, vlhics@rit.edu

## ABSTRACT
Non-visual text-entry for people with visual impairments has focused mostly on the comparison of input techniques reporting on performance measures, such as accuracy and speed. While researchers have been able to establish that non-visual input is slow and error prone, there is little understanding on how to improve it. To develop a richer characterization of typing performance, we conducted a longitudinal study with five novice blind users. For eight weeks, we collected *in-situ* usage data and conducted weekly laboratory assessment sessions. This paper presents a thorough analysis of typing performance that goes beyond traditional aggregated measures of text-entry and reports on character-level errors and touch measures. Our findings show that users improve over time, even though it is at a slow rate (0.3 WPM per week). Substitutions are the most common type of error and have a significant impact on entry rates. In addition to text input data, we analyzed touch behaviors, looking at touch contact points, exploration movements, and lift positions. We provide insights on *why* and *how* performance improvements and errors occur. Finally, we derive some implications that should inform the design of future virtual keyboards for non-visual input.

## Categories and Subject Descriptors
H.5.2 [**Information Interfaces and Presentation**]: User Interfaces - *Input devices and strategies*. K4.2 [**Computers and Society**]: Social Issues – *Assistive technologies for persons with disabilities*.

## General Terms
Measurement, Experimentation, Human Factors.

## Keywords
Blind, Novice, Text-Entry, Input, Touch, Behavior, Performance.

## 1. INTRODUCTION
Over the last decade, touchscreen devices began to dominate the smartphone market. In contrast to feature phones, current devices are operated by touching the screen directly, without requiring a physical keyboard; users resort to virtual keyboards to enter text on their devices. Although text-entry is an inherently visually demanding task, particularly when using touchscreen devices, accessibility services have been devised to enable blind users to perform this task[1].

These services rely on an *Explore by Touch* paradigm where users move their fingers on the screen and the interface reads aloud the element in focus. While *Explore by Touch* can be useful, the fundamental task of text input remains slow and error prone, especially for novice users [1, 3, 19]. Although touch interaction and non-visual text-entry have been studied for years, research has been mainly limited to performance comparisons of input techniques [1, 3, 19, 22, 28]. In these studies, performance is often measured in terms of words per minute and errors. While these measures can establish *that* differences exist, they provide little justification for *why* and *how* they exist. In order to improve current input techniques, an understanding of touch typing behaviors is essential. However, because there is little or no knowledge on how blind users type on standard virtual keyboards it is unclear how to improve them.

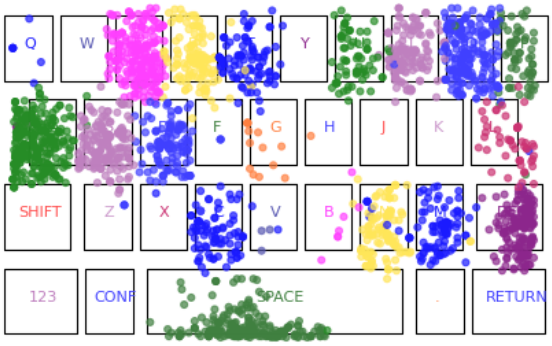

**Figure 1. Lift points for all participants in week eight.**

This paper presents a longitudinal study that aims to understand the characteristics of novice blind users typing performance, touch behaviors (e.g. exploration gestures - Figure 1), character-level errors, and learning experience. Our goal is to inform future designs of touchscreen keyboards, with the ultimate aim of supporting fast and accurate input that can be easily used by novice blind users. We recruited five blind participants and gave them new mobile touchscreen devices. A service that ran in the background of the devices collected system-wide input usage, which enabled us to account for practice. We then collected typing data through controlled laboratory assessments over an eight-week period. To develop a detailed characterization of how blind users explore and type on virtual keyboards, we propose extending text-entry measures to include movement measures, as in pointing-related research [9, 10, 12], such as distance traveled, target re-entries, and movement profile. These measures can provide more insights on how blind people use a continuous interaction paradigm such as *Explore by Touch*. We were interested in answering three main research questions: 1) What are the most common types of errors? 2) How does typing performance evolve over time? 3) *Why* do errors exist?

Our findings have implications for the design of touchscreen keyboards and input techniques for blind and visually impaired users. Based on typing data, our results show that substitutions are the most common error type throughout the study. Participants' performance significantly improved over time, both in terms of errors and speed. We also show why improvements occur by examining hit positions, movement time, movement paths, and

[1] https://www.apple.com/accessibility/ios/voiceover/

pausing behaviors. Correction strategies were consistent among users, but required a significant amount of time.

The main contribution of this paper is a thorough understanding of unconstrained text-entry performance, typing behaviors, and an empirical body of knowledge for future development of virtual keyboards. We provide an analysis of touch exploration measures in text-entry tasks and report on the learning experience of novice blind users, particularly on how input performance and behaviors change over an eight-week period. The findings herein presented should be of interest to mobile keyboard designers and accessibility researchers looking to gain from quantitative insights into blind users' text-entry performance with touch devices.

# 2. TEXT INPUT FOR BLIND USERS

Today's mainstream touchscreen devices support non-visual text input via the built-in screen readers e.g. VoiceOver and Talkback. They enable users to explore the keyboard with their finger and have the keys are read aloud as they touch them. While the visual layout of the QWERTY keyboard is identical to that presented to sighted users, the text-entry rates are much slower for visually impaired users [19]. To address this problem a number of works have proposed novel interfaces for non-visual text-entry on mobile touchscreen devices; including new keyboard layouts [3, 8] and alternative methods of inputting text [1, 16, 19, 22, 23].

What is common amongst these works is that they focus on the overall input performance metrics such as words per minute (WPM) and minimum string distance (MSD) error rates [21] to compare input methods. However, in doing so these works neglect to justify *how* and *why differences* exist between interfaces; for instance, in character-level errors [13, 27]. Similarly, the aforementioned studies use constrained text-entry tasks – where the participants are not provided with feedback on their input actions, or given the ability to correct errors. In contrast, Wobbrock and Myers [27] presented the input stream taxonomy to support unconstrained text-entry evaluations. This approach allows participants to make corrections to their typing and capture both uncorrected and corrected error rates. Using this analysis, it is possible to not only capture character-level errors, but also identify corrective behaviors.

# 3. MEASURES OF TOUCH BEHAVIORS

Findlater et al. [5] evaluated the typing performances of expert sighted typists on large touch surfaces. Through an analysis of touchscreen measures, they identified individual differences in key centroids and hit point deviations (i.e. x and y offsets of touch gestures with regards to individual keys). Later, they proposed personalized keyboards that could adapt to individual typing patterns and improve entry rates [6]. Guerreiro et al. [7] applied similar touch measures to investigate tablet text-entry behaviors of blind users with one- and two-handed input. While the text input performance metrics revealed no statistical difference between conditions, using the x, y offsets of the initial touch down positions, the authors uncovered that users landed closer to intended keys with two-handed input. Furthermore, when measuring movement distances of non-visual exploration, participants using two hands performed more efficient paths through the keyboard. The authors leveraged the fact that non-visual touchscreen interactions result in gestures with periods of continuous movement and traces through the interface, opposed to the discrete point interactions of sighted users.

While using movement measures is uncommon when analyzing text input, they are well established within cursor movement research. MacKenzie et al. [12], proposed seven accuracy measurements to understand users' behaviors with pointing devices. Included in these were path analysis measurements, such as target re-entries, task axis crossing, movement direction and orthogonal direction change. The authors also proposed continuous measures such as movement variability, errors and offsets. Hwang et al. [9] believed analysis of submovements within pointing device selections could reveal new insights into the challenges faced by motor-impaired users. To understand individual differences between motor-impaired users' cursor movements Hwang et al. proposed analyzing the number and duration of pauses, verification times, submovements within the intended target, target slips, and velocity profile of movements.

In this paper, we extend on existing text-input analysis techniques and propose the inclusion of discrete and continuous touch movement measurements to better understand touchscreen text input behaviors of blind users.

# 4. LONGITUDINAL USER STUDY

Prior research investigating non-visual text-entry on mobile devices has merely reported on the overall text input performance measurements, failing to examine the underlying characteristics of users' typing behaviors. We believe that a detailed analysis of text-input, using the proposed touch measurements, are key to expose the challenges faced by novice blind users. Our ultimate goal is to identify new opportunities to reduce the learning overhead and support better non-visual input on mobile touchscreen devices. In order to achieve these goals, we conducted a longitudinal study with blind novice smartphone users. Participants were each provided with a mobile device preloaded with our data collection tool and asked to use the device as their primary phone for eight weeks. Due to the ethically sensitive nature of the research, no participants were asked to consent to their data being shared beyond the research group and as such supporting data cannot be made openly available.

## 4.1 Participants

We recruited five participants with visual impairments, four males and one female, from a local training institution for blind people. Participants' age ranged from 23 to 55 (M=37.2, SD=15.2) years old, and all participants were legally blind as defined within our IRB approved recruitment criteria. They were experienced desktop screen reader users. However, none owned a smartphone or had prior experience with touchscreen screen readers.

## 4.2 Procedure

Our study was designed to capture the progression of typing performance of novice users. Prior laboratory studies of longitudinal text-entry evaluations report using seven sessions with noticeable improvements [1]. Thus, we decided that eight weeks (weekly sessions) would be sufficient to observe comparable progression. The user study consisted of two components: in-situ usage and weekly laboratory assessments.

### 4.2.1 In-Situ Device Usage

Our goal was to collect everyday text-entry usage by novice blind users. Participants received basic training on how to use a virtual keyboard. We did not define, incentivize or force usage protocols. Instead, we developed a data collection framework that ran as a background service on their smartphones and collected usage measures, i.e. *time spent using applications* and *number of keystrokes entered*. Previously, Evans and Wobbrock [4] demonstrated that it is possible to obtain text-entry performance measurements (speed and errors) from everyday computer usage. However, analysis of everyday mobile typing performance is out with the scope of this paper – in-situ data will be used to give

context of the device usage for individual participants and support the analysis of our weekly laboratory text-entry evaluations.

#### 4.2.2 Weekly Lab Text-Entry Evaluations

Participants met with the researchers weekly, for eight weeks, and performed 20 minutes of text-entry trials. Each trial contained one sentence comprised of five words, with an average size of 5 characters, and a minimum correlation with language of 0.97. We developed an experimental application that would select the trial sentences from a written language corpus. The application randomly selected the sentences for the session to avoid order effects and captured transcribed sentences and completion times. The experimental application started the trial by reading the target sentence aloud via the device's TTS (Text-to-Speech engine). Upon finishing each sentence, participants pressed the return key twice to advance to the next trial. They were encouraged to type as accurately and quickly as possible. We used an unconstrained text-entry protocol [27], where participants were free to correct any errors they encountered. To ensure that participants would not practice the trials outside the laboratory evaluations, the application was installed on the participants' device at the beginning of each session, and uninstalled at the end. Automatic correction and cursor movement operations were not used during the trials.

Our study was carried out in Portuguese, as such there are a number of letters that are uncommon in the written language, and therefore do not appear within our trial sentences (e.g. W and Y). Subsequently, these keys will contain no examples of intended interactions within our evaluation.

### 4.3 Apparatus

Participants were each provided with a *Samsung S3 Mini* touchscreen smartphone, running Android 4.1 operating system. We enabled the *Talkback* screen reader and pre-installed our data collection service, TinyBlackBox (TBB). TBB was designed to constantly run in the background, capturing users' interactions with the device. This approach enabled us to capture text-entry usage data throughout the eight-week period.

The S3 Mini default input method was Samsung's own Android QWERTY keyboard. Although visually the keys have both horizontal and vertical spacing, when *Talkback* is enabled and the participants touch the screen, they receive feedback for the nearest key to their touch point. However, when moving from one key to another, the key with current focus occupies the spacing. This means that target boundaries can grow and shrink based on the exploration paths. S3 Mini's default keyboard was used throughout our study, both in laboratory evaluations and in-situ.

### 4.4 Dependent Measures

Text-entry performance was measured by analyzing trials' input stream [27]. We report on words per minute (WPM), total error rates, uncorrected error rates, and corrected error rates. Moreover, we investigate character-level errors and types of errors (substitutions – incorrect characters, insertions – added characters, and omissions – omitted characters). Touch exploration behaviors were measured using x, y positions and variability [5] (hit point deviations), movement time, movement distances, Path Length to Task Axis length ratio (PL/TA), count and duration of pauses within the movements [9, 10, 12], and visited keys.

### 4.5 Design and Analysis

We performed Shapiro-Wilk tests on all dependent measures. For normally distributed values we used a mixed-effects model analysis of variance [15]. Mixed-effects models extend repeated measures models, such as ANOVAs, to allow unequal number of repetitions; that is, unbalance data such as ours, where we have different numbers of trials per week for each participant. We modeled *Week* as a fixed effect and *Trial* was included as a nested factor within *Week*. In addition, *Participant* and the interaction between *Participant* and *In-Situ Usage Time* were modeled as random effects to account for correlated measurements within subjects over time [24].

For the measures that were not normally distributed, we applied $\log_e$ or $\log_{10}$ transforms [2], which resulted in normally distributed measures [Shapiro-Wilk $p>.05$]. We then used the mixed-effects model terms previously described for further analysis.

## 5. RESULTS

Our goal is to characterize novice blind users' text-entry performance and learning when using *Explore by Touch*. We describe participants' in-situ usage and relate it with text-entry performance. We analyze input speed, accuracy, and character-level errors over an eight-week period. Finally, we characterize users' touch exploration behaviors and provide insights on *how* and *why* input performance changes over time.

### 5.1 In-Situ Usage

Table 1 and Table 2 summarize the number of characters entered and time spent typing, respectively.

**Table 1. Characters entered in-situ. Columns represent weeks.**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| **P1** | 245 | 405 | 555 | 678 | 799 | 133 | 732 | 1292 |
| **P2** | 1283 | 648 | 1548 | 5396 | 1248 | 411 | 2120 | 208 |
| **P3** | 75 | 697 | 579 | 1115 | 310 | 1205 | 1 | 447 |
| **P4** | 1002 | 1022 | 566 | 601 | 2435 | 603 | 2578 | 1099 |
| **P5** | 32 | 45 | 22 | 21 | 12 | 24 | 189 | 383 |

**Table 2. Time spent typing in-situ (minutes).**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| **P1** | 66.2 | 62 | 46.6 | 54.6 | 101 | 26.7 | 46.5 | 85.9 |
| **P2** | 180 | 53.6 | 98.7 | 383 | 92.8 | 29.8 | 149 | 12.3 |
| **P3** | 1.78 | 85.8 | 99.1 | 170 | 40.7 | 131 | 0 | 57.7 |
| **P4** | 160 | 196 | 43 | 36.5 | 127 | 36.5 | 201 | 91 |
| **P5** | 5.25 | 3.7 | 7.4 | 1.5 | 0.45 | 1.17 | 15.2 | 65.3 |

Participants entered a total of 32,764 characters over eight weeks. They spent a total of 51 hours actively entering text. Generally, the number of characters entered is directly related with time spent. However, there is a high variance in usage results both between participants and weeks. For instance, while P2 and P3 were particularly active in the fourth week, others such as P4 were more active in the last two weeks. P5 was the least active with an average usage of 12.5 minutes (SD=20) per week. On the other hand P2 and P4 spent on average 125 (SD=110) and 111 (SD=65) minutes typing per week. Although analysis of in-situ performance measures is out of the scope of this paper, we will leverage usage data to control for performance improvements in all statistical analysis.

### 5.2 Text-Entry Performance

In total, participants produced 11,560 characters from which 1,323 were backspaces, resulting in 10,237 transcribed characters. In this section we thoroughly analyze input performance regarding speed and accuracy over an eight-week period.

#### *5.2.1 Input Speed*

To assess input speed, we used the words per minute (WPM) measure calculated as (length of transcribed text – 1) * (60 seconds / trial time in seconds) / (5 characters per word).

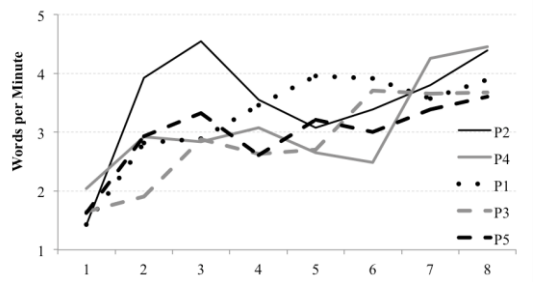


**Figure 2. Words per minute over 8 weeks.**

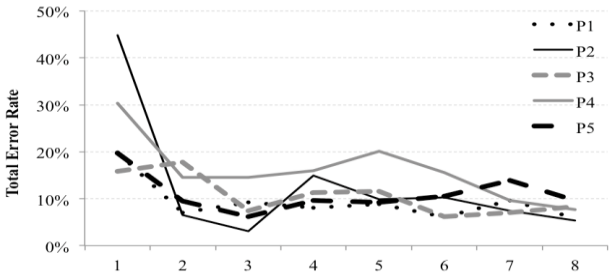


**Figure 3. Total error rate (%) over 8 weeks.**

***Slow learning rate.*** Participants improved on average 2.4 wpm (SD=.36) from week one with 1.6 wpm (SD=.23) to 4 wpm (SD=.35) after eight weeks. We found a significant effect of *Week* on *WPM* [$F_{1,7}$=12.329, $p$<.001] as all participants improved over time. Nevertheless, considering that participants were familiar with QWERTY keyboards, learning rates are still low with an average improvement of 0.3 wpm per week.

***Still improving after eight weeks.*** Figure 2 shows WPM graphed over eight weeks. We can see that participants are still improving input speeds at the end of the user study. Fitting power laws [25] to entry rates and extrapolating to twice the weeks gives an average entry speed of 5 wpm in week 16th.

***External factors can negatively influence performance.*** We can also notice that P2 and P4 have atypical changes in performance in week four and seven, respectively. When debriefing P2 about this sudden drop in performance, she mentioned perceiving the speech feedback being slower while typing after installing a 3rd party app, *WhatsApp*[2]. In fact, this is a known issue with this particular application. Although we are not able to confirm that speech feedback changed, we can show that both number of pauses and duration of pauses during movement, increased from week 3 to week 5, while movement speed and distance traveled decreased in the same time period (see Section 5.4). This suggests that external factors had an influence in this participant's typing behavior (e.g., other apps or emotional issues).

***In-situ usage improves performance.*** Regarding P4, the abrupt increase in input speed is most likely related with the increase of usage in week seven (see Tables Table 1 and Table 2). After debriefing P4 in that week, he mentioned that he was finally using his phone to the fullest, particularly sending and receiving text messages. He stated "*... the phone is finally fully accessible to me, I can send SMS, I can send text messages via Skype, I can send all the messages that I want*". Therefore, we believe the sudden increase in input speed is due to his increase in usage of messaging applications. In fact, we found a significant medium size effect between *Input Speed* and *In-Situ Usage time* [Pearson's $r_{(290)}$=.353, $p$<.001].

### *5.2.2 Input Accuracy*

In order to analyze input accuracy, we calculated: 1) uncorrected - erroneous characters in the final transcribed sentence, 2) corrected - erased characters that were erroneous, and 3) total error rates - erroneous characters that were entered (even those that were corrected) [27].

***Total error rates tend to 7.4%.*** P2 achieved the highest total error rate of 45% on week 1 and finished the user study with the lowest rate of 5.4% by week 8. Overall participants started with an average total error rate of 26% (SD=11.7%) and finished with 7.4% (SD=1.7%) [$F_{1,7}$=4.176, $p$<.001]. Moreover, Figure 3 shows that error rates start to stabilize around that value.

***Errors are usually corrected.*** Table 3 shows the uncorrected error rates for each participants and week. Overall, when given the chance, users tend to correct most errors, resulting in high quality transcribed sentences. This goes in line with previous findings for sighted users [21]. For instance, P1 and P2 had the lowest uncorrected error rates with 0% and 0.3% by week 8. On average, participants left only 1.6% (SD=1.4%) errors in the transcribed sentences by week 8, which resulted in a significant effect of *Week* [$F_{1,7}$=2.306, $p$<.05].

**Table 3. Unc. error rates (%). Columns represent weeks.**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| **P1** | 4 | 0.4 | 1.9 | 1.4 | 2.3 | 0.3 | 2.6 | 0 |
| **P2** | 1 | 1 | 0.3 | 0 | 0 | 0 | 1.5 | 0.3 |
| **P3** | 7.6 | 8.5 | 3.4 | 4.1 | 0.5 | 2.8 | 1.9 | 2.5 |
| **P4** | 20 | 4.7 | 5.2 | 6.3 | 7.8 | 3.2 | 3.2 | 1.9 |
| **P5** | 11 | 5.6 | 4.3 | 5.3 | 5.3 | 2.3 | 5.1 | 3.3 |

**Table 4. Corrected error rate (%). Higher is better.**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| **P1** | 74 | 77 | 63 | 89 | 81 | 81 | 77 | 91 |
| **P2** | 87 | 55 | 73 | 89 | 84 | 91 | 85 | 68 |
| **P3** | 62 | 50 | 41 | 72 | 50 | 46 | 71 | 57 |
| **P4** | 69 | 81 | 69 | 68 | 71 | 56 | 62 | 60 |
| **P5** | 86 | 100 | 60 | 50 | 92 | 86 | 89 | 88 |

***23-39% of deletions were inefficient.*** Corrected error rates illustrates the amount of effective *"fixing"* and allows to answer the question "of the erased characters, what percentage were erroneous?" High rate means that most of erased characters were errors and should have been corrected. Participants achieved average corrected error rates between 61% (SD=12%, week 3) and 77% (SD=11%, week 7), which means that 23% to 39% of deleted characters had been correctly entered. This occurs because errors are not immediately recognized. For instance, when phonetically similar characters are entered (e.g. N→M), users only notice that mistake when the word is read aloud. To fix the error, several characters, including correct characters, are usually deleted. A detailed inspection of logs files shows that editing operations, such as cursor movement, were never used. Average corrected error rate per week is 73%, which remains fairly constant throughout the eight weeks [$F_{1,7}$=.98, $p$=.447].

***13% of time is spent correcting errors.*** The time spent correcting errors is subsumed by input speed (see Section 5.2.1); however, such analysis does not provide insights on the cost of such corrections. Examining correcting actions shows that participants spent on average 32% (SD=17%, MIN=19% [P5], MAX=65% [P2]) of their time correcting text in the first week. Performance significantly improved over time and by week eight only 13% (SD=1.8%) of time was spent in this task [$F_{1,7}$=4.806, $p$<.001].

[2] https://www.whatsapp.com/

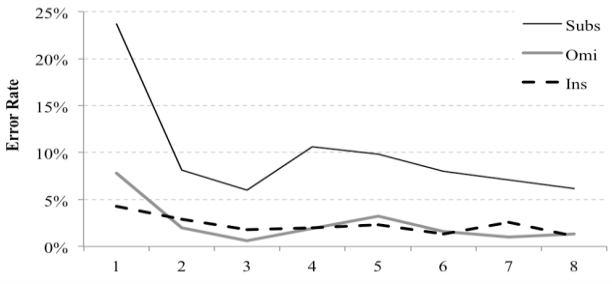


**Figure 4. Types of error over 8 weeks.**

## 5.3 Character-Level Errors

In this section, we present a fine grained analysis by categorizing types of input errors: insertions, substitutions, and omissions [14]. We report aggregate measures, which represent the method's accuracy over all entered characters, but also at the level of individual letters [27]. These findings can help designers in addressing specific types of errors and characters.

***Substitutions are the most common type of error.*** Figure 4 illustrates the types of errors over the eight-week period. Substitution errors were consistently higher than insertions and omissions. Although there was a significant decrease in substitution error rates over time, from 24% (SD=12%) to 6% (SD=1%) [$F_{1,7}$=3.518, $p$<.005], they still remain significantly higher than the remaining types of errors [$F_{2,8}$=125.321, $p$<.001]. In fact, substitution error rate is higher than omissions and insertions combined. This result holds true for all participants.

***Similar substitution rates across keys.*** Overall, participants had similar error rates across all intended keys. No row, column, or side patterns emerged from weekly data. Moreover, keys near edges had similar accuracy rates to those in the center (Figure 5).

***No clear substitution pattern.*** To analyze the most common substitution errors, we created confusion matrices. In week eight, some of the most common substitutions were Q→E (33%), B→H (17%), P→O (9%), P→L (4%), R→T (4%). Unlike sighted users that experience substitution patterns towards a predominant direction [5, 17], blind users' patterns are less clear. This is most likely related with the differences between visual and auditory feedback when acquiring keys. Further discussion on this topic is available in Section 5.4.

***Adjacent phonetically similar characters promote substitutions.*** Since feedback is solely auditory, phonetically similar characters have the potential to be confused when blind users are exploring the keyboard. In the Portuguese language, particularly when using Android's Text-to-Speech engine, there are three cases prone to confusions: I-E, O-U, and M-N. For I-E substitution error rates are constantly low over time (0-1%) and inexistent from week five. Regarding O-U substitutions, error rates are slightly higher with 8.5% in week one and decreasing to 0.5% in week eight. Finally, concerning M-N substitutions, error rates remain between 3% (week one-three) and 6.5% (week five) across the eight-week time period. Indeed, in week eight, error rates are still 4.5%. No other adjacent pair of letters obtained such a consistently high (and symmetrical) error rate over time. These results suggest that phonetically similar letters that are close together have higher probability of being substituted.

***68% of omission errors are left uncorrected.*** Omission error rates decreased 6.5% from week one (M=8% SD=6%) to week eight (M=1% SD=0.7%) [$F_{1,7}$=3.858, $p$<.005]. Unlike substitutions, the majority of omission errors are not corrected. On average 68% (SD=14%) of errors are left uncorrected. These errors are usually described as cognitive errors [11]. A common explanation is misspellings or users forgetting to type certain letters. However, leaving errors uncorrected may also be related with (lack of) feedback after an attempt to enter a character, confirming that an input action had a consequence. This option seems less likely since users received feedback after each character entry. Although omissions only account for 2.4% of errors (see Figure 4), they are the least likely to be corrected.

## 5.4 Touch Exploration Behaviors

In this section we provide new insights on participants' touch exploration behaviors. We examine the three stages that compose a key selection: touching the screen, moving the finger to find the intended key, and lifting the finger. For this analysis, we removed outlying points where the entered key (on lift) was more than one key distance away from the intended key in either *x* or *y* direction to account for transposition or misspelling errors.

### *5.4.1 Hit Point Analysis*

Hit points correspond to landing positions. It is noteworthy that at this point, users do not have any feedback about the key they will land on. Unlike input from sighted users, which aims towards visual stimuli, blind users solely resort to their spatial model of the keyboard and some physical affordances (e.g. device size).

***Users land on intended keys nearly half the times.*** By week eight, 48% (SD=12%) of key presses landed within the boundaries of intended targets. This number may seem low, but it is not unexpected given that participants did not receive any auditory feedback until this point. Nevertheless, performance significant increased from week one (M=27%, SD=15) to week eight [$F_{7,28}$=5.222, $p$<.01], showing that users gain a better spatial model of the keyboard. We found that at week eight, 91% (SD=5%) of the times, participants land either inside the intended key or an adjacent key. Also, landing on the correct row (M=78%, SD=7%) is easier than landing on the correct column (M=59%, SD=11%) [$F_{1,4}$=27.611, $p$<.01], which is not surprising given that rows make larger targets than columns.

**Keys near physical edges are easier hit.** Throughout the eight-week period, keys that are positioned on physical edges are easier to land on. For instance, in week eight, participants correctly landed on characters A and Q in 75% and 71% of times, respectively. On the other hand, characters such as B or M were only correctly hit 14% and 16%, respectively. The space bar consistently outperforms the remaining keys (week eight M=99%), most likely due to a combination of its positioning (on the bottom edge) and width (five times larger).

***Emergent keyboard is shifted towards the bottom and most key overlaps are horizontal.*** We examined the emerging key shapes and sizes using hit points. Figure 6 illustrates the emergent keyboard for week eight; that is, the keyboard layout that results from participants' touches. In week one, the key sizes are larger and shifted towards the center of the screen, where users started their exploration, which resulted in larger overlaps between keys.

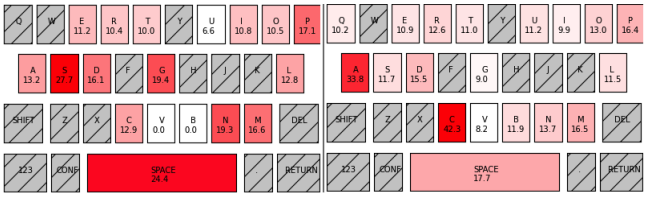


**Figure 5. Substitution error rates per key. Gray keys were not used in the trials. Darker colors indicate higher error rates. Left - week 2, Right - week 8.**

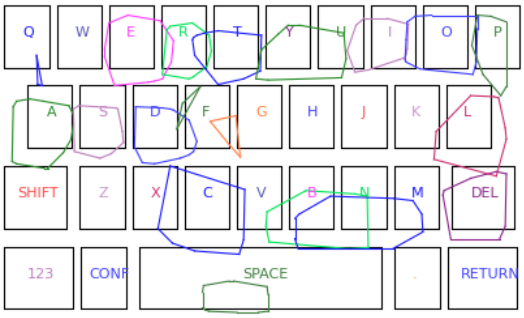


**Figure 6. Polygons encompass hit points within a standard deviation of key centroid.**

By week eight, participants are able to land nearer to keys; however, there are still significant overlaps, mostly horizontally. Characters M and N are particularly interesting, since they present the largest overlap (Figure 6). Also, we can see that hit points tend to occur below the center of the intended target.

### 5.4.2 *Movement Analysis*

Previous research has investigated text-entry performance by blind users. However, analyses tend to focus on performance measures, such as time and errors. In this section, we aim to establish *why* performance improvements occur by conducting a through analysis of touch exploration behaviors.

***Users visit on average one extra key.*** In the first week, the average number of visited keys per keystroke was 4.9 (SD=1.9). Participants significantly improved their performance achieving an average of 2 visited keys (SD=0.3) by week eight [$F_{7,28}$=5.133, $p$<.001]. Similarly, the number of target re-entries (entering the same target for the second time) also improved from 6.6 (SD=3.2) to 0.8 (SD=0.3) [$F_{7,28}$=7.498, $p$<.001]. This corresponds to an average of 49 traveled pixels (SD=11), where 60% of movement is done in the x-axis, which is consistent with previous results where users are more likely to land on the intended row and then perform horizontal movements.

***Users learn how to perform more efficient explorations.*** In order to understand exploration efficiency, we calculated the Path Length (movement distance) to Task Axis length (Euclidean distance between hit point and center of target) ratio. Participants significantly improved over time from 3.6 (SD=1.3) to 0.95 (SD=0.15) [$F_{7,28}$=6.033, $p$<.001]. Notice that we obtained an average ratio below 1 because the Task Axis length is the distance to the center of the target. Users only require traveling to the edge of the target in order to select the key.

***Keystroke time is on average 1.9 seconds.*** In line with previous touch measures, movement times also improved from 4.1 seconds (SD=1.4) to 1.9 seconds (SD=0.3) [$F_{7,28}$=5.424, $p$<.001]. This value may seem high, but it is expected since users need to wait for auditory feedback to confirm which letter they are touching. As a consequence, entry times are directly related to speech rate and delay. Figure 7 illustrates P1's dwell times in week one and eight. Longer pauses are clearly visible in the first week. Also, because feedback is received when entering keys, pauses often occur near their edges.

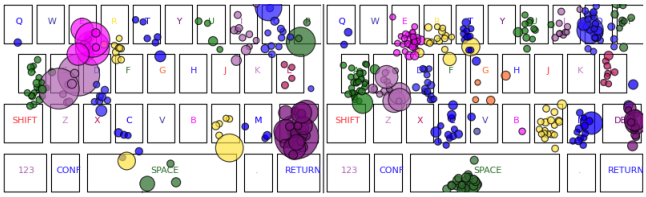

**Figure 7. A circle indicates a pause; size represents its duration. Left - week 1 for P1, Right - week 8 for P1.**

***Keys near physical edges require less time to press but do not result in lower error rates.*** We found significant differences between keys located near the device's edge, such as Q, A, P, and L, and all other keys regarding movement time [week eight, Z=2.032, $p$<.05]. Nevertheless, this difference does not result in accuracy improvements. In fact, border keys have a slightly higher substitution rate (week eight, 7% vs. 5.4%, n.s.).

***Insertion errors have smaller movement times and distances.*** Insertion errors are related to unintentionally and accidentally entered characters. Knowing how to filter these keystrokes can result in performance improvements. When analyzing movement times and distances, we found significant differences between correct entries and insertion errors [$F_{1,4}$=23.287, $p$<.01; $F_{1,4}$=24.119, $p$<.01] throughout the eight-week period. These results suggest that touch data can be used to classify insertions.

### 5.4.3 *Lift Point Analysis*

Where hit point and movement analysis examined where users land on the screen and how they explore the keyboard, respectively, an examination of lift point allows us to understand the final step of selecting a key. It is particularly relevant to understand in what conditions substitution errors occur.

***Lift points are spread-out over keys' boundaries.*** Figure 1 illustrates all lift points for week eight. Data shows that points are spread over intended keys and particularly close to their edges. Unlike sighted users [5, 17, 18], there is not a clear touch offset direction, which can have significant implications when building touch models for this user group. Moreover, hit point deviations (standard deviations) remain unchanged across time with 25.6px in week one and 24.3px in week eight, which is approximately half the size of a key. This suggests that users may be prone to slip errors; that is, slipping to a nearby key just before selecting it.

***There is more to substitution errors than slips.*** We classified as finger slips all entries where the last visited key was the intended target. Although we are not applying a time threshold, this measure gives us all entries that need to be considered as slip errors. Overall, in week one 37.5% (SD=17%) of substitution errors were slips. In week eight we obtained a similar value of 38.4% (SD=12%) [$F_{1,7}$=2.095, $p$>.05]. Notice that slip errors account for less than 50% of substitution errors by week eight. Taking into account that users should receive speech feedback before selecting the intended key, we analyzed whether participants' finger paths crossed it at some point during movement. In week eight, for 64% (SD=9.8%) of substitution errors, participants were inside the boundaries of the target at some point in their touch paths; however, failed to select it in a timely manner. After identifying some of the instances where these errors occurred, we conducted a manual examination of the recorded videos. We noticed that most of the cases were related to a significant delay between speech feedback, which resulted in a mismatch between the key being heard and touched at that moment. Participants tried to compensate for this delay by performing corrective movements, but often resulted in entering the incorrect key. Further research should explore this issue by investigating the effect of auditory delay on input accuracy.

***For some substitutions, intended keys are not even visited.*** According to the results described above, in week eight there are still 36% of substitutions where participants do not even visited the key they were aiming for. This means that they performed a selection without hearing the intended key. From visual inspection

of individual keystrokes' movements, we derived several reasons for this behavior: *1) Accidental touches* – similarly to insertion errors, participants unintentionally touch the keyboard close to the intended character. These keystrokes are short in distance and time. *2) Phonetically similar keys* – this happens when users cross a key that sounds similar to the intended character (e.g. while aiming for M, the user lands on B, moves to the right, enters N, and lifts the finger), resulting in a substitution error. *3) Overconfidence on spatial model* – in some substitution instances it seems that participants overly rely on their spatial understanding of the keyboard by performing a gesture and selecting a key without waiting for feedback. Lastly, *4) Feeling lost and giving up* – some exploration paths show fine-grain movements near the intended key, going back and forth; however, participants never hit the intended character.

# 6. DISCUSSION

In this section we describe major results, implications for future design of virtual keyboards, and limitations of our work.

**Summary of Major Results**

Participants achieve an *average typing speed* of 4 WPM and 4.7% total error rate after eight weeks of usage. Although performance was still improving in the last week, *learning rate was slow* (0.3 WPM per week). Previous research has shown similar results when analyzing overall typing performance [1]. An open question until now was: why and how did users improved typing performance? Overall participants seem to gain a better spatial model of the keyboard by *landing closer to targets*, performing more *time- and movement-efficient paths* towards intended targets, and *less target re-entries*, which resulted in lower number of pauses to hear auditory feedback.

Character-level analysis revealed that most erroneous characters are *substitutions*. However, in contrast with sighted typing patterns, results do not show a clear offset pattern. Instead, touch *points are scattered over intended keys* and particularly near edges. Substitution errors can have different causes and *slip errors* only account for about 38% of these cases. One would assume that participants would only lift their fingers once they hear the intended key; however, by week eight, this is not the case for 36% of substitutions errors.

Finally, participants naturally correct the overwhelming majority of errors (98.4%), which corresponds to about *13% of their typing time*. Moreover, one third of corrections are *counterproductive* as users delete correct characters.

**Implications for Design**

*Easier, effective, and efficient correction.* Corrections are still time consuming and inefficient. None of our participants used cursor-positioning operations throughout the study. It seems that these actions are only expected to be used by expert typists, preventing novice users to do fine-grain corrections. Also, participants did not use auto-correct or auto-complete solutions, although these have great potential to be used in non-visual text-entry to correct missed errors (such as omissions) and improve typing speeds.

*Synchronize speech output with touch input.* Results suggest that 64% of substitution errors can be due to a mismatch between speech output and touch information. Future non-visual keyboards should prioritize synchronization between input and output modalities.

*Filter unintentionally added characters.* Accidental touches originate substitution and insertion errors, which in turn take time to correct. However, most of these errors can be filtered out by monitoring movement's time and distance, since they are significantly shorter than correct entries.

*Use language-based solutions.* The majority of omission errors (68%) go by undetected and therefore uncorrected. Language-based solutions such as spellcheckers seem to be the only plausible solution. Nevertheless, mainstream auto-correct approaches should also be able to deal with some substitution errors. Current algorithms usually weight word corrections by keyboard distance. Although blind users do not show a predominant touch offset direction, most substitution errors were adjacent keys.

*Leverage land-on and movement information.* Non-visual typing comprises much more than just lift positions. Movement data can provide evidence of what particular key users are trying to select. Future key recognizers should leverage this information and try to predict the most probable targets (see [20, 26] for pointing prediction). This information could be used with language models to narrow the search space of word-corrections or provide character suggestions when users delete a letter.

*Touch models need to adapt to expertise.* Leveraging movement data is particularly relevant on early stages of learning when users perform longer exploration paths. While expert users may land on the intended target most of the times, novice users still need to search for the intended key and wait for auditory feedback. Therefore, touch models need to be able to adapt to different typing behaviors (i.e. abilities) and learning rates.

**Limitations**

Our participants only included five novice blind users. Although this is a small number of participants they represent a crucial user group when the goal is to designing easy-to-use solutions and identify challenges with current virtual keyboards. Although typing performance and touch behaviors can be significantly different for expert users, the derived implications may still apply. For instance, using more efficient correction strategies or language-based solutions can further improve experts' typing performance. Further research should replicate the analysis reported in the paper with more experienced blind typists in order to examine character-level errors and touch movement behaviors.

Finally, in this user study participants were allowed to use their device in-the-wild. Although we were able to control for device usage in our analysis, our weekly laboratory assessments may have influenced learning results. Thus, it is likely that reported weekly performance may not represent a truly natural learning experience; however, it surely represents the challenges users face while learning to type on virtual keyboards.

# 7. CONCLUSION AND FUTURE WORK

We have investigated the unconstrained typing performance and touch exploration behaviors of 5 novice blind users over the course of an eight-week period. Results show that users improve both entry speed and accuracy, although at slow rate. Improvements are mostly due to a combination of factors, such as landing closer to intended keys, performing more efficient keyboard explorations, lower number of target re-entries, and lower movement times. By week eight, users land inside the intended key or adjacent keys 91% of the time. The most common error type is a substitution. Regarding correction strategies, users correct most of typing errors, which consumes on average 13% of input time. Overall, we provide a thorough examination on how blind users type using a virtual keyboard. Future work should

apply the design implications that emerged from our results and develop new solutions to improve typing performance.

## 8. ACKNOWLEDGMENTS

We thank all participants from Fundação Raquel e Martin Sain in Lisbon. This work was partially supported by: RIT's Center for Accessibility and Inclusion Research, RCUK (EP/G066019/1), and FCT (SFRH/BD/103935/2014, UID/CEC/00408/2013).

## 9. REFERENCES

[1] Azenkot S. et al. 2012. Input Finger Detection for nonvisual touch screen text entry in Perkinput. *Proceedings of Graphics Interface (GI '12)* (2012).

[2] Berry, D.A. 1987. Logarithmic transformations in ANOVA. *Biometrics*. (1987), 439–456.

[3] Bonner, M. et al. 2010. No-Look Notes: Accessible eyes-free multi-touch text entry. *Pervasive Computing*. (2010), 409–426.

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

[5] Findlater, L. et al. 2011. Typing on flat glass: examining ten-finger expert typing patterns on touch surfaces. *Proceedings of the 2011 annual conference on Human factors in computing systems* (2011), 2453–2462.

[6] Findlater, L. and Wobbrock, J.O. 2012. Personalized Input: Improving Ten-Finger Touchscreen Typing through Automatic Adaptation. *Proceedings of the 2012 annual conference on human factors in computing systems (CHI'12)*. (2012).

[7] Guerreiro, J. et al. 2015. TabLETS Get Physical: Non-Visual Text Entry on Tablet Devices. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2015).

[8] Guerreiro, T. et al. 2008. From tapping to touching: Making touch screens accessible to blind users. *IEEE MultiMedia*. (2008), 48–50.

[9] Hwang, F. et al. 2004. Mouse movements of motion-impaired users: a submovement analysis. *Proceedings of the 6th international ACM SIGACCESS conference on Computers and accessibility* (New York, NY, USA, 2004), 102–109.

[10] Keates, S. and Trewin, S. 2005. Effect of age and Parkinson's disease on cursor positioning using a mouse. *Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility* (2005), 68–75.

[11] Kristensson, P.O. 2009. Five challenges for intelligent text entry methods. *AI Magazine*. 30, 4 (2009), 85.

[12] MacKenzie, I.S. et al. 2001. Accuracy measures for evaluating computer pointing devices. *Proceedings of the SIGCHI conference on Human factors in computing systems* (2001), 9–16.

[13] MacKenzie, I.S. and Soukoreff, R.W. 2002. A character-level error analysis technique for evaluating text entry methods. *Proceedings of the second Nordic conference on Human-computer interaction* (2002), 243–246.

[14] MacKenzie, I.S. and Soukoreff, R.W. 2002. Text entry for mobile computing: Models and methods, theory and practice. *Human--Computer Interaction*. 17, 2 (2002), 147–198.

[15] McCulloch, C.E. and Neuhaus, J.M. 2001. *Generalized linear mixed models*. Wiley Online Library.

[16] Nicolau, H. et al. 2014. B#: Chord-based Correction for Multitouch Braille Input. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2014), 1705–1708.

[17] Nicolau, H. and Jorge, J. 2012. Elderly Text-Entry Performance on Touchscreens. *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility* (2012).

[18] Nicolau, H. and Jorge, J. 2012. Touch typing using thumbs: understanding the effect of mobility and hand posture. *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems* (New York, NY, USA, 2012), 2683–2686.

[19] Oliveira, J. et al. 2011. Blind people and mobile touch-based text-entry: acknowledging the need for different flavors. *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility* (2011), 179–186.

[20] Pasqual, P.T. and Wobbrock, J.O. 2014. Mouse pointing endpoint prediction using kinematic template matching. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (2014), 743–752.

[21] Soukoreff, R.W. and MacKenzie, I.S. 2003. Metrics for text entry research: an evaluation of MSD and KSPC, and a new unified error metric. *Proceedings of the SIGCHI conference on Human factors in computing systems* (2003), 113–120.

[22] Southern, C. et al. 2012. An Evaluation of BrailleTouch: Mobile Touchscreen Text Entry for the Visually Impaired. *Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services* (New York, NY, USA, 2012), 317–326.

[23] Tinwala, H. and MacKenzie, I.S. 2010. Eyes-free text entry with error correction on touchscreen mobile devices. *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (2010), 511–520.

[24] Verbeke, G. and Molenberghs, G. 2009. *Linear mixed models for longitudinal data*. Springer Science & Business Media.

[25] Wobbrock, J.O. 2007. Text Entry Systems: Mobility, Accessibility, Universality. I.S. MacKenzie and K. Tanaka-Ishii, eds. San Francisco: Morgan Kaufmann. 47–74.

[26] Wobbrock, J.O. et al. 2009. The angle mouse: target-agnostic dynamic gain adjustment based on angular deviation. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2009), 1401–1410.

[27] Wobbrock, J.O. and Myers, B.A. 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.

[28] Yfantidis, G. and Evreinov, G. 2006. Adaptive blind interaction technique for touchscreens. *Universal Access in the Information Society*. 4, 4 (2006), 328–337.