Highlights

The Guided Evaluation Method: An Easier Way to Empirically Estimate Trained User Performance for Unfamiliar Keyboard Layouts

Aunnoy K Mutasim, Anil Ufuk Batmaz, Moaaz Hudhud Mughrabi, Wolfgang Stuerzlinger

- Repeated word/phrase typing has limitations in estimating trained user performance
- A novel Guided Evaluation Method (GEM)
- The GEM can estimate trained user text entry performance within minutes

The Guided Evaluation Method: An Easier Way to Empirically Estimate Trained User Performance for Unfamiliar Keyboard Layouts

Aunnoy K Mutasim[a,*], Anil Ufuk Batmaz[b], Moaaz Hudhud Mughrabi[c], Wolfgang Stuerzlinger[a]

[a] *School of Interactive Arts & Technology (SIAT), Simon Fraser University, Vancouver, BC, Canada*
[b] *Computer Science & Software Engineering Department, Concordia University, Montreal, QC, Canada*
[c] *Mechatronics Engineering Department, Kadir Has University, Istanbul, Turkey*

Abstract

To determine in a user study whether proposed keyboard layouts, such as OPTI, can surpass QWERTY in performance, extended training through longitudinal studies is crucial. However, addressing the challenge of creating trained users presents a logistical bottleneck. A common alternative involves having participants type the same word or phrase repeatedly. We conducted two separate studies to investigate this alternative. The findings reveal that both approaches, repeatedly typing words or phrases, have limitations in accurately estimating trained user performance. Thus, we propose the Guided Evaluation Method (GEM), a novel approach to *quickly* estimate trained user performance with novices. Our results reveal that in a matter of minutes, participants exhibited performance similar to an existing longitudinal study — OPTI outperforms QWERTY. As it eliminates the need for resource-intensive longitudinal studies, our new GEM thus enables much faster estimation of trained user performance. This outcome will potentially reignite research on better text entry methods.

Keywords: Text Entry, Touch Typing, OPTI, QWERTY, GEM, Soft Keyboards

*Corresponding author

1. Introduction

Text entry is today an essential component of people's everyday life and related activities range from texting, sending emails, writing reports and other documents, to an alternative means of communication for people with limited muscle control [1, 2, 3, 4]. Currently, most text is entered either via physical or soft keyboards. Although several more or less optimal layouts have been presented in the literature [5], the QWERTY keyboard layout and its variations continue to be the primary means of inputting text to this day for languages that use the Latin-script alphabet [6, 7].

The main reason behind this predominance is the wide availability of QWERTY keyboards for more than a century. Thus, many people have been trained on this layout and it is habitually used in their everyday lives [7]. The text entry speed achievable via QWERTY is not only very familiar to users but is also considered to be the most appropriate baseline for anyone experimenting with unfamiliar keyboard layouts.

Previous research has shown that alternative keyboard layouts that are theoretically more efficient than QWERTY can offer significant advantages, but only after substantial training [8]. For example, users needed an average of 4 hours of practice with a new keyboard layout called OPTI to achieve the same typing speed as with QWERTY [8]. In the context of touch-based typing [5, 8], OPTI has been predicted to be approximately 30% faster than QWERTY, but this speed advantage is typically only achieved after a total of 17 hours of training. Such long training periods make it difficult to persuade regular computer and smartphone users to commit to such training, especially when the current typing speed offered by QWERTY is already deemed to be satisfactory for most situations.

Yet, for input techniques that are more challenging to use, like eye-gaze-based pointing or brain–computer interfaces for individuals with limited muscle control [1, 2, 3, 4], optimal keyboard layouts have the potential for a proportionally higher impact on typing speeds and may therefore justify the effort required to learn a new layout. Still, as previously mentioned, evaluating the performance potential of unfamiliar layouts suffers a major methodological bottleneck. Demonstrating the performance potential of a given layout requires novices (i.e., first-time or beginner-level users of a system) to undergo some form of training that can range from about 30 minutes per participant per layout to up to 7.5 hours of each participant's time over 14 days, or similar time commitments, e.g., [8, 9, 10, 11, 12, 13, 14, 15, 16,

17, 18, 19, 20, 21, 22, 23, 24, 25]. Thus, conducting studies that require participants to be extensively trained is logistically expensive, especially when comparing novel or unfamiliar layouts to QWERTY [8, 21], even more so when comparing multiple different layouts within the same study.

To investigate whether it is possible to speed up the evaluation process of unfamiliar layouts, we first define a more precise meaning of a few terms. In the text entry literature, participants who have been trained over a few sessions on the same day or via a typical text entry longitudinal study have been (loosely) referred to as "experts", e.g., [17, 26, 27, 28, 29]. We disagree with this definition and instead support the notion that to become an expert one needs to train for years, or go through approximately 10,000 hours or more practice [30]. Thus, we refer to longitudinal study participants [4, 8, 16, 17, 18, 19, 20, 21, 22] or users who have been trained for a time frame that is comparable to such studies, simply as "*trained users*". Similarly, we refer to participants who have undergone training only through multiple sessions on the same day [6, 9, 10, 11, 12, 13, 14, 15] as "*minimally trained users*".

To reduce the logistical burden of a longitudinal study, some researchers have used an approach that requires participants to only type the same word(s)/phrase(s) repeatedly, e.g., [14, 21, 27, 28, 31]. Yet, this approach for estimating trained user performance through repeatedly typing the same word/phrase has not been validated before. More specifically, it is still unknown *if repeatedly typing the same word/phrase provides a good estimate of trained user text entry performance* for a given keyboard layout. To address this gap in the literature, we first investigate whether this approach provides a good estimate of a layout's potential performance with trained users. Towards this goal, and similar to previous work [8], we compare OPTI and QWERTY by tasking participants to type the same word/phrase repeatedly.

The outcomes of a first repeated phrase typing study show that this approach does not accurately estimate trained user-level text entry performance within a single day's training. Still, this approach can potentially produce good estimates, but only through a longitudinal study that takes multiple days [32]. We then investigate the approach of repeated word typing. When we analyze the data word-by-word, i.e., weighting all words the same regardless of their length, the results came close to that of previous work [8]. Yet, we note that this is a biased outcome, as word length distributions are not uniform in human languages, and analyzing the data word-by-word thus does not appropriately estimate real keyboard usage. Once we take the word length into account in the analysis, the results are far from what was reported

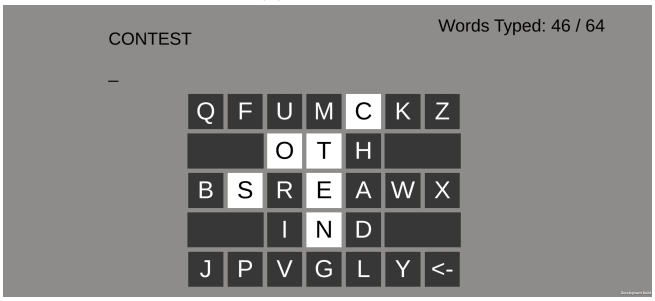
in the literature [8].

Building on these outcomes, we investigate a different approach to bypass the need for longitudinal studies. One core difference between novices and trained/expert users in terms of their keyboard usage is that trained/expert users are able to locate keys more quickly than novices, incurring substantially less average visual search time [33, 34]. To reduce the visual search time, and in the process, address the lengthy time requirements for longitudinal studies, we introduce here our novel **Guided Evaluation Method (GEM)** that allows researchers to empirically evaluate the performance of trained users in a more efficient manner. The GEM simply highlights all the keys in a target text that are yet to be typed. Also, instead of presenting a whole phrase, the GEM tasks participants to type individual words and requires them to first plan the typing pattern of that word before actually entering it (see Figure 1). All these features combined significantly reduce the visual search time of novice users and contribute towards making such users perform like trained users. Yet, although the GEM potentially simulates the typing behavior of a trained user, we acknowledge that the GEM involves a typing experience that is not typical for real-life typing. Thus, we emphasize that the GEM is intended solely as an evaluation method to quickly assess trained user performance and/or for the comparison of different keyboard layouts.

To investigate whether this novel evaluation approach estimates the performance of a trained user correctly, we replicate here an existing longitudinal study with the GEM [8]. Beyond assessing two other evaluation methods for text entry systems, our main question is *whether the GEM approach can estimate trained user performance correctly and if the GEM reduces the effort during the evaluation of text entry systems in research on novel keyboard layouts.*

Our main contributions include: 1) showing that the popular approach of repeatedly typing the same *phrase* does not accurately estimate trained user-level text entry performance within a single day's training, 2) demonstrating that the repeated *word* typing approach is limited in terms of validity and reliability if analyzed at the word level, and does not yield accurate estimates when the word length distribution is accounted for in the analysis, and, 3) proposing the GEM approach and demonstrating that it can accurately estimate trained user-level performance for text entry studies in a *matter of minutes.*

(a) QWERTY



(b) OPTI

Figure 1: The two keyboard layouts used in our three user studies, where only Study 3 involved the Guided Evaluation Method (GEM) with key highlighting. The GEM highlights all the keys in the target word that are yet to be typed in white. With the GEM, participants are also asked to first plan the typing pattern of each target word before they start typing and then to type the word as quickly as possible. In (a), as the letter 'L' has already been typed and, as it does not exist elsewhere in the remainder of the target word, its corresponding key is no longer highlighted.

2. Literature Review

2.1. Evaluation Methods for Text Entry Systems

Previous text entry studies typically trained novices from 30 minutes to up to 7.5 hours per participant per layout to evaluate the performance of a given text entry method/layout. This training was either conducted on the same day, e.g., [6, 9, 10, 11, 12, 13, 14, 15], or, following a more externally valid approach, over several days through a longitudinal study, e.g., [4, 8, 16, 17, 18, 19, 20, 21, 22]. Kjærup et al.'s work [35] provides a comprehensive literature review of such longitudinal studies.

MacKenzie and Zhang [8] compared the QWERTY keyboard to the theoretically more optimal OPTI layout in such a longitudinal study. There, five participants used a stylus to type on each of the two keyboards. A single session consisted of two 20-22 minute long typing rounds, one for each layout. Each of the 20 sessions per participant was separated by an interval ranging from two hours to two days. The results showed that the average typing speed for OPTI increased from 17.0 words per minute (WPM) in the first session to 44.3 WPM in the 20^{th} one, while QWERTY started at about 27.5 WPM and ended at about 40 WPM in the final session. The participants were able to type faster with OPTI after about 4 hours of practice, i.e., starting from the 11^{th} session. However, according to MacKenzie and Zhang [8], even though participants had about 7 hours of practice with OPTI across the 20 sessions, i.e., they were trained users, this surely does not make them true experts. To address this, the authors used the power law of learning [4, 18, 19, 21, 36] to extrapolate the data up to the 50^{th} session and suggested Equations 1 and 2 for QWERTY and OPTI, respectively.

$$WPM_{QWERTY} = 27.597 \times session^{0.1237}, R^2 = 0.9802 \quad (1)$$

$$WPM_{OPTI} = 17.24 \times session^{0.3219}, R^2 = 0.9974 \quad (2)$$

In Equations 1 and 2, R^2 is the squared correlation coefficient and $session$ corresponds to the number of 20-22 minute training sessions required. According to these equations, after 17 hours of practice, QWERTY and OPTI could potentially reach 44.8 WPM and 60.7 WPM, respectively. Thus, for touch-based keyboards, MacKenzie and Zhang showed that theoretically more optimal keyboard layouts can outperform the more popular QWERTY layout [8], but only after *lengthy* training.

In a different effort, and to improve the speed of dwell-based systems for gaze-based keyboards, Majaranta et al. [16] conducted a longitudinal study where the participants were allowed to adjust the dwell time required to select a key according to their preference. Eleven participants typed on a QWERTY keyboard on ten separate days with each session/day lasting 15 minutes. The results showed that their typing speed improved from 6.9 WPM during the first session to 19.9 WPM in the last. The average dwell time decreased from an average of 876 ms to 282 ms in the 10^{th} session. Still, this study took again substantial time, specifically, ten days per participant, to perform.

To reduce the logistical burden of such a longitudinal study, researchers have experimented with an approach that only requires participants to repeatedly type the same word(s)/phrase(s), e.g., [14, 21, 27, 28, 31]. However, typing just one phrase does not accurately reflect the frequency of characters in the language, i.e., English in our context. Therefore, this approach cannot reliably evaluate real keyboard usage. To address this, Yu et al. [14] used a slightly modified approach where each participant was given a unique phrase to transcribe 12 times. A further improvement was presented in a later work [37], where the participants were tasked with typing the same *ten* phrases in each session for eight sessions.

However, as speculated by Jokinen et al. [18], while repeatedly typing the same word on an unfamiliar keyboard layout *"may"* demonstrate performance that outperforms QWERTY, typing the same *phrase* over and over may not exhibit the same results. Recently, we verified Jokinen et al.'s speculation that the approach of repeatedly typing the same phrase on an unfamiliar layout like OPTI does not outperform QWERTY within a single day's training [32]. That work also indicated that this approach is able to provide a good estimate of trained user performance and can do so faster than traditional training with different phrases [32]. Yet, doing this still requires training participants over several days, which is still equivalent to a longitudinal study, albeit a comparatively shorter one.

Mathematical models (e.g., [5, 18, 33, 36, 38, 39]) and empirical studies relying on a system that simulates a perfect recognizer [9] have also been used to predict the performance potential of a given layout. In the context of touch-based text entry, Rick [5] developed a model that predicted that swipe-based typing (e.g., [40]) has the potential to achieve faster text entry speeds compared to tap-based typing. For example, a 17.3% gain over tapping is expected for QWERTY. Yet, if a "more suitable" layout is used, e.g., OPTI

II, a much higher typing speed could be achieved. More specifically, OPTI II is predicted to enable typing at a 29.5% faster rate than QWERTY.

Magnien et al. [41] introduced another approach that provides visual clues to novice participants. Once the user has inputted a few characters, the system highlights the next few probable characters in bold, which reduces the visual search space. Their results demonstrated that this approach could significantly decrease the time required to enter text. A similar study was conducted recently by Grüneis et al. [42], too. They showed that a visual clue presented either as a pop-up animation or with an enlarged font size achieved the best performance, and improved the typing speed by approximately 36% and 34%, respectively. In our GEM approach, we build on the idea of providing visual clues [41, 42] to reduce the visual search space. However, the methods proposed in previous work [41, 42] do not replicate the actions of a trained/expert user since users still need to visually search for the next key (albeit among a comparatively smaller subset of keys). We mitigate the issue of the required visual search time to some extent in the GEM approach by requiring participants to first also *plan* their typing of a word, before actually starting to enter the word.

2.2. Touch-Based Text Entry

Previous work has presented multiple novel methods for typing text on touch-based devices. This includes but is not limited to: hand-posture adaptive keyboards [43, 44], key-target resizing keyboards [45], tap-stroke hybrid keyboards [46], gesture keyboards [47, 48, 49, 50, 51], and different keyboard layouts [20, 52, 53, 54]. However, the focus of such studies is on improving typing speeds or exploring new text input methods. In contrast, our objective is not to develop an improved touch-based text entry system. Instead, we take advantage of the ubiquity of touch-based systems to validate the GEM by building on MacKenzie and Zhang's [8] work. Yet, with the GEM, we are presenting a methodology that can accelerate the search for better text entry layouts and/or techniques that support this area of research.

3. User Study 1 – Phrase Repetition

Our first study [32] was designed to evaluate the approach of typing the same *phrase* repeatedly for the two keyboard layouts used by MacKenzie and Zhang [8], OPTI and QWERTY. We chose (a slightly modified version of) the method used in MacKenzie and Zhang's study and its results

[8] as a comparison point for two reasons. First, this choice allowed us to compare our results with the findings of an independently performed and thus more externally valid longitudinal study. Also, unlike, e.g., work on T9 [55], this study utilizes a form of touch-based input, which matches today's smartphone-based text entry methods reasonably well.

3.1. Apparatus

To ensure that participants were using a device that they were familiar and comfortable with from their everyday use and to make it also easier to recruit remote participants, we asked them to perform the experimental task using their personal smartphone.

3.2. Keyboard Designs

To ensure comparability to MacKenzie and Zhang's work [8], we aimed to replicate their work as closely as possible. Therefore, we deliberately did not consider typing disambiguation methods, such as tap positions [5] or language models [40], as including these could have confounded the results.

In our study, participants typed on both the OPTI and QWERTY keyboard layouts with the device held in landscape mode. To ensure both layouts fit comfortably onto the screen, we limited the study to smartphones with a display size of at least 6" diagonally. However, to accommodate the keyboards onto a 6" screen, we had to reduce the key sizes by 0.2 cm compared to the original design by MacKenzie and Zhang [8], resulting in keys that were 0.8×0.8 cm. This slight decrease in key size should not significantly impact the results, as suggested by Fitts' law [56, 57, 58, 59, 60] and previous research [29, 33]. We also introduced a gap of 0.1 cm between keys to prevent unintended selections when users touched the edge between two keys. Additionally, we provided auditory feedback in the form of a subtle click for each key press. We also replaced the F1 key in the original OPTI layout with a backspace key to facilitate error correction. Following typical smartphone QWERTY layouts, the backspace key was added next to the character 'M'. We showed participants their task progress in the top-right corner of the screen. The design of the two keyboard layouts is illustrated in Figure 1. We developed the Android app for the study using Unity.

3.3. Participants

Eight participants (7 male), aged 31.4 ± 5.21 years, took part in the study. They were recruited via word of mouth or ads over social media platforms

and email and were compensated $15 for their participation. All participants had over nine years of experience with typing on a physical/soft QWERTY keyboard, but none had experience with the OPTI keyboard prior to this study.

3.4. Procedure

Participants started by filling out a consent form and a demographic questionnaire, including questions regarding their age, gender, and experience using OPTI and QWERTY layouts. Then, they typed using both OPTI and QWERTY, presented in counterbalanced order in a within-subjects experimental design with a single independent variable — the two keyboard layouts. Each participant typed the *same* phrase 96 times over eight sessions (i.e., 12 times per session) for each of the two layouts. Participants were instructed to type an extra "space" after the last word, to also denote the end of that phrase [24]. Participants were asked to rest for 2 minutes between sessions.

The phrases were randomly selected from MacKenzie and Soukoreff's set [61], which consists of a total of 500 phrases. The minimum, maximum, and average phrase lengths in that set are 16, 43, and 28.6 characters, respectively, and the average word length is 4.45 characters. To increase the external validity, we followed Yu et al.'s approach [14] and ensured that every participant was given a different/unique phrase to transcribe.

Instead of the stylus used by MacKenzie and Zhang [8], participants were instructed to use their dominant hand's index finger [18] to type on the mobile screen. They were also instructed to correct any mistakes immediately if they noticed them but to ignore mistakes that occurred two or more letters back. We made this choice to avoid the cumulative effect of too many corrections, which in the process can substantially affect the typing speed [62]. Participants performed the task while sitting in a chair and holding the phone with their non-dominant hand. Participants were asked to rest for at least 5 minutes between the two keyboard conditions.

At the end of the experiment, we conducted a short semi-structured interview where participants were asked to share thoughts about their typing experience with the two layouts. The experiment took about an hour per participant, including the demographic questionnaire and the interview.

3.5. Performance Metrics

In this study, we chose the following metrics to evaluate the text entry performance of the two keyboard layouts:

- *Words per minute (WPM)* represents the number of words typed per minute. Here, a single word is defined as a sequence of any 5 characters. This includes spaces but excludes backspaces [63].

- *Keystrokes per character (KSPC)* is the average number of key selections required to (correctly) type in a single character. More precisely, KSPC is the ratio of the total number of selected keys to the length of the typed text [64]. Thus, KSPC takes into account the number of times the backspace key was hit.

- *Minimum String Distance Error Rate (MSD ER)*, where the MSD represents the minimum amount of changes, including insertions, deletions, and substitutions, needed to convert one string to another. We use the MSD ER metric formulation proposed by Soukoreff and MacKenzie [65] to compute the difference between the target and the entered phrase.

Table 1: WPM results for each session, overall totaling about 25 minutes of typing on each keyboard. Also, projections via Equations 1 and 2 of how many (22 minutes long) sessions or how much time would be required to reach the corresponding WPM with the traditional approach of typing different phrases during training.

	Typing Speed (WPM)		**Projected No. of Training Sessions**		**Projected Training Time (hours)**	
Session	**OPTI**	**QWERTY**	**OPTI**	**QWERTY**	**OPTI**	**QWERTY**
1	19.1 ± 5.85	33.7 ± 6.37	1.4	5.0	0.50	1.83
2	22.9 ± 7.01	36.0 ± 6.35	2.4	8.6	0.89	3.14
3	25.4 ± 7.11	35.1 ± 6.12	3.3	7.0	1.23	2.56
4	25.7 ± 7.24	36.7 ± 6.15	3.5	10.0	1.27	3.67
5	27.9 ± 8.08	35.5 ± 5.76	4.5	7.6	1.64	2.80
6	31.0 ± 7.97	34.4 ± 7.73	6.2	5.9	2.28	2.16
7	30.8 ± 7.86	34.1 ± 5.6	6.1	5.5	2.22	2.03
8	31.0 ± 8.64	33.0 ± 5.94	6.2	4.3	2.27	1.57

3.6. Results and Discussion

Our results are presented in Table 1 and Figure 2. As per Figure 2a, although repeatedly typing the same phrase improved participants' performance for OPTI, the speed for OPTI never approached the performance achieved with QWERTY. In other words, participants' typing speed using the QWERTY layout was always above the typing speed achieved using OPTI throughout the eight sessions.

As can be seen in Table 1 and Figure 2a, participants' average typing speed with OPTI was 19.1 ± 5.85 WPM in the first and 31.0 ± 8.64 WPM in the last session. The speed achieved in the last session was also the fastest average typing speed across participants in that condition. The fastest individual participant reached an average speed of 41.1 ± 5.12 WPM in the 8^{th} session with OPTI. Using QWERTY, participants achieved on average 33.7 ± 6.37 WPM in the first and 33.0 ± 5.94 WPM in the last session. The fastest average typing speed of 36.7 ± 6.15 WPM was achieved by the participants on the 4^{th} session with QWERTY (see Table 1). The fastest individual participant was able to reach an average typing speed of 42.6 ± 7.87 WPM on the 4^{th} session as well with the QWERTY layout.

For both OPTI and QWERTY layouts, no noteworthy trends were observed in terms of KSPC and MSD ER. Other than a few exceptions, i.e., in sessions 1 and 3-5 in Figure 2b, and sessions 3 and 8 in Figure 2c, the curves for both keyboards look fairly similar to each other. Also, the average MSD ER for both keyboard layouts was below 2% for all sessions (see Figure 2c), showing that the participants were quite careful when using both keyboard layouts [40].

When referencing our results with Equations 1 and 2, which are based on MacKenzie and Zhang's work [8], the average typing speed of the last session is equivalent to 4.3 (20-22 minute long) sessions for QWERTY and 6.2 sessions for OPTI (see Table 1). For QWERTY, where the fastest session was the 4^{th} session, we estimate that it would take 10 sessions to reach that speed of 36.7 WPM. This shows that repeatedly typing the same phrase for about 30 minutes can – at best – achieve comparable results to 3.67 hours of normal/traditional training (i.e., training with different phrases) with a known keyboard layout, i.e., QWERTY, or 2.27 hours of traditional training with an unknown layout, in this case, OPTI.

In contrast to our results, MacKenzie and Zhang [8] found OPTI was able to outperform QWERTY starting from the 11^{th} (20-22 minute) session and eventually reached 44.3 WPM in the 20^{th} session, with QWERTY achieving

about 40 WPM on the same session. The typing speed of our participants did not even get close to the performance of the trained users in MacKenzie and Zhang's study [8] for either keyboard. The results of our study thus indicate that – although repeatedly typing the same phrase improves performance – this approach is not suitable for reliably estimating trained user performance for unfamiliar keyboard layouts. Thus, our results support Jokinen et al.'s [18] speculation that repeatedly typing the same phrase on an unfamiliar layout, e.g., OPTI, does not demonstrate that it can surpass QWERTY's performance, at least not within a single day's training. More importantly, *the approach of repeatedly typing the same phrase does not yield a good estimate of trained user-level text entry performance with a single day's training*.

The change in typing speed over time for OPTI in Figure 2a showed an increasing trend until the 6^{th} session. From there onwards, the curve flattened out completely. We see this as evidence that continuing the experiment for a few more sessions probably would not have increased OPTI's performance any further. As for QWERTY's trend over time in Figure 2a, the typing speed shows a slight increase from the 1^{st} to the 4^{th} session. However, a decreasing trend can be observed from the 5^{th} session onwards.

We probed this issue in our semi-structured interviews, where one participant mentioned that *"the task very quickly got boring and frustrating as it seemed like it was never going to end."* Another shared *"My mind kept wandering off. It was very hard to continuously keep my concentration on the task."* Similarly, another participant explained *"After a couple of sessions, QWERTY was especially hard as there was no challenge associated with the task, ... , unlike OPTI, where I felt there was still scope for improvement."* Others also gave similar feedback about the experimental task. Given these insights, the downward trend of QWERTY's typing speed starting from the 5^{th} session is most likely associated with participants' fatigue and the lack of challenge in the experimental task. This brings up the question of whether this fatigue had an effect on the results for the flattening trend of OPTI starting from the 6^{th} session.

To investigate this issue, we fit a regression based on the power law of learning through the WPM data [8, 18, 19, 21, 36], see Figure 3, which yielded the following two equations:

$$WPM_{QWERTY} = 35.684 \times session^{-0.016}, R^2 = 0.0517 \tag{3}$$

$$WPM_{OPTI} = 15.64 \times session^{0.3273}, R^2 = 0.9596 \tag{4}$$

In Equations 3 and 4, a single *session* comprises repeatedly typing the same phrase 12 times, and R^2 represents the squared correlation coefficient. According to the R^2 value in Equation 4, approximately 96% of the variance is accounted for in the fitted learning model, which means that the model predicts user behaviors very well. Equation 4 predicts that it would require ≈24 sessions (289 repetitions, 8 sessions per day = a little over 3 days) of repeatedly typing the same phrase with OPTI before participants would reach the typing speed of 44.3 WPM, which was reported by MacKenzie and Zhang [8] after twenty (20-22 minute) normal/traditional sessions of training, i.e., training with different phrases. In other words, although repeatedly typing the same phrase can *eventually* provide a good estimate of trained user performance and can do so faster than traditional training, this approach still requires training participants over several days and therefore does **not** eliminate the need for a longitudinal study.

Yet, typing the same phrase repeatedly on a known layout, in this case, QWERTY, had detrimental effects on learning. This is evident by the (small) negative exponent (i.e., -0.016) and a very small R^2 value in Equation 3. Thus, we suggest not to use the repeatedly typing the same phrase task for more than 4 sessions, i.e., 48 repetitions, for QWERTY or any other layout that is already highly familiar to a participant. Overall, our analysis based on the power law of learning showed that the observed flattening trend of OPTI's typing speed from 6^{th} session and the decreasing trend of QWERTY from 5^{th} session onwards are most probably due to the fatigue associated with the experimental task of typing the same phrase repeatedly.

The repetitive phrase typing approach has been used in the past either with complete novices [31], or with participants who first trained for a few sessions using the traditional approach of typing different phrases and then typing the same phrase for a few more sessions, e.g., [14, 21, 27, 28]. In our study, we chose the former method [31] which uses a single phrase repetitively. We did this as we wanted to investigate an approach that has a higher probability of quickly reaching trained user performance. Further, we also ensured that the total number of repetitions for a single phrase was substantially larger than the combination of phrases and repetitions typically employed by other studies [14, 21, 28].

4. User Study 2 – Word Repetition

The results from Study 1 identify that the repeated phrase typing method is not a good substitute for longitudinal studies, specifically within a single day of training. Thus, in the second study presented here, we investigated the approach of repeatedly typing the same *word* [28], a method that Jokinen et al. [18] postulated to be potentially more promising compared to repeated phrase typing.

For this study, we (again) conducted a modified replication of MacKenzie and Zhang's study [8], with the difference of employing the repeated word typing approach. This choice also permits us to validate the effectiveness of the repeated word typing approach relative to previous work. Here, we outline the study's specifics and the methodology used to assess the performance of the two keyboard layouts, OPTI and QWERTY.

4.1. Participants

7 females and 7 males, i.e., 14 participants aged 22.7 $\pm$ 3.24 years, took part in the word repetition study. 10 participants had experience with typing on a physical and/or soft QWERTY keyboard for over nine years, two participants had been typing for 7-9 years, and another two for 5-7 years. All participants typed on the OPTI keyboard for the first time in this study. They received a compensation of $15 for their participation.

4.2. Apparatus, Keyboard Designs, and Procedure

The same apparatus and keyboard design from Study 1 were employed in this study. Participants started by completing a consent form and a demographic questionnaire covering information such as their age, gender, and familiarity with the OPTI and QWERTY layouts. Subsequently, participants proceeded to type six phrases (randomly chosen from MacKenzie and Soukoreff's set [61]) on each of the two keyboards presented in counterbalanced order. In this study, the phrases were presented to them one word at a time, and each such word was presented repeatedly. Unlike work that used the approach of typing a random word (or the same phrase) repeatedly [14, 27, 28, 31, 32], our choice of using the approach of presenting phrases word-by-word still ensures that the character frequency, sequence, and length of the typed content closely resemble real keyboard usage and thus our results should match many other existing text entry studies, e.g., [8, 16, 40], as closely as possible.

Once each target word had been entered, we asked participants to type a "space" to clearly denote the end of that word, like in Study 1. The reason why we chose "space" to end each word is not only because that is how words typically end, but also because it is the most frequently typed key [8, 66]. For each word, participants were tasked to repeatedly type that word 10 times, after which the next word in the current/next phrase appeared. At the end of repeatedly typing a word 10 times, the keyboard disappeared for 200 ms and then reappeared again to give participants a visual cue that the target word had changed. We also asked them to memorize each target word, including its spelling, if necessary.

The participants did not undergo practice trials; instead, we treated all ten repetitions of all the words of the first phrase as practice, i.e., the whole first phrase, and excluded that data from the analysis [14, 67]. Participants were asked to rest for at least 5 minutes between the two keyboard layouts. After completing the experiment, participants filled out a brief questionnaire, sharing their preferences and providing feedback on aspects such as ease of interaction, frustration, mental, and physical fatigue, as well as perceived speed and precision for each keyboard, using 7-point Likert scales. On average, participants spent approximately 20 minutes typing on each keyboard. The entire experiment, encompassing the demographic pre-questionnaire and post-questionnaires, took about an hour.

For all other aspects, we adhered to the same procedure as in Study 1, which involved restricting touch typing to the index finger of the dominant hand and focusing correction efforts solely on mistakes within the last two letters.

4.3. Performance Metrics

We used the same performance metrics as in Study 1, i.e., WPM, KSPC, and MSD ER, here as well. However, we used two different approaches to analyze the data:

- *Word-level* — With this approach, we calculated the results word-by-word for every repetition, similar to existing word repetition studies [28].

- *Phrase-level* — In this approach, we combined the words, which were presented separately to the participants, *across* each repetition into the target phrase. So, all third repetitions of each word of the phrase

were "assembled" into the third instance of the whole phrase. Then, we calculated the results for the entire phrases. This approach matches how text entry results are typically analyzed in most studies, e.g., [6, 8, 9, 10, 11, 12, 13, 15, 16, 17, 20, 21, 23, 24, 25, 28, 68, 69], and thus enables comparisons of our results with a wide range of work in the literature.

More importantly, in the phrase-level approach, combining the words into phrases ensures that the word length distribution of a given language, i.e., English in our case, is also taken into account in the calculation of the results. In other words, the word-level approach considers every word, irrespective of its length, to have the same weight in the calculation of the average WPM/KSPC/MSD ER. In contrast, the phrase-level approach weighs each word by its length, preserving the word length distribution in its calculation, thus computing an appropriately weighted average. With our approach, we also address a limitation of a previous word repetition study [28], where the average and maximum word lengths of the chosen 20-word set were just 3.3 and 5, respectively, whereas the average word length is 5.1 in the English language [70], which means that this previous study may have biased results towards shorter words (that are easier to enter).

4.4. Experimental Design

We used a within-subjects design for our evaluation, with two independent variables — the keyboard layout and the repetition number. As dependent variables, we measured participants' WPM, KSPC, and MSD ER. For each of these performance metrics, we calculated the word-level and phrase-level averages for each of the 10 word-by-word repetitions of the five typed phrases, i.e., ignoring the first practice phrase. On average, each participant typed 27.2 words ten times for each layout, with the average word length being 5.38.

4.5. Results

We first present an analysis of the main objective measures followed by the subjective ones. The objective measures were analyzed using two-way repeated measures (RM) ANOVA with $\alpha = 0.05$ in SPSS 29. The subjective measures data was analyzed using dependent t-tests with $\alpha = 0.05$ using the same software. We considered data to be normally distributed when Skewness (S) and Kurtosis (K) values were within ± 1.5 [71, 72]. For RM

ANOVA, upon violation of Mauchly's sphericity test, we applied Huynh-Feldt correction where $\epsilon < 0.75$, and all post-hoc analyses were conducted with the Bonferroni method. For dependent variables that did not have a normal/log-normal distribution, data was transformed using Aligned Rank Transform (ART) [73]. For brevity, we detail only statistically significant results.

Table 2: Word-level RM ANOVA results for the two Keyboard Layouts and ten Repetitions.

	Keyboards	Repetitions	Keyboards × Repetitions
WPM	$F_{1,13} = 0.88, n.s.,$ $\eta^2 = 0.064$	$\boldsymbol{F_{3.80,49.4} = 50.8, p < 0.001,}$ $\boldsymbol{\eta^2 = 0.796}$	$\boldsymbol{F_{3.53,45.9} = 53.0, p < 0.001,}$ $\boldsymbol{\eta^2 = 0.803}$
MSD ER	$\boldsymbol{F_{1,13} = 8.38, p < 0.05,}$ $\boldsymbol{\eta^2 = 0.392}$	$\boldsymbol{F_{9,117} = 2.25, p < 0.05,}$ $\boldsymbol{\eta^2 = 0.147}$	$F_{9,117} = 1.78, n.s.,$ $\eta^2 = 0.121$

Table 3: Word-level WPM results for each repetition, overall totaling about 20 minutes of typing on each keyboard. Also, projections via Equations 1 and 2 of how many (22 minutes long) sessions or how much time would be required to reach the corresponding WPM with the traditional approach of typing different phrases during training.

	Typing Speed (WPM)		**Projected No. of Training Sessions**		**Projected Training Time (hours)**	
Repetition	**OPTI**	**QWERTY**	**OPTI**	**QWERTY**	**OPTI**	**QWERTY**
1	23.6 ± 15.5	36.3 ± 10.8	2.6	9.1	0.97	3.34
2	37.3 ± 16.9	38.6 ± 10.3	11.0	15.0	4.05	5.50
3	39.5 ± 15.8	38.1 ± 10.2	13.1	13.5	4.82	4.95
4	40.3 ± 15.5	38.0 ± 10.3	14.0	13.2	5.12	4.85
5	40.5 ± 14.7	37.5 ± 9.54	14.2	12.0	5.21	4.39
6	41.3 ± 14.5	37.8 ± 9.42	15.1	12.7	5.54	4.65
7	41.4 ± 14.2	38.1 ± 9.70	15.2	13.6	5.57	4.98
8	40.9 ± 14.3	37.8 ± 8.99	14.7	12.7	5.38	4.65
9	42.0 ± 13.2	37.6 ± 9.24	15.9	12.3	5.84	4.51
10	42.4 ± 13.4	37.6 ± 8.88	16.4	12.1	6.02	4.44

*4.5.1. **Word-level Analysis** of WPM, KSPC, MSD ER, and Required Number of Traditional Training Sessions*

All significant results for our main object measures are presented in Tables 2 and 3, and Figures 4 and 5. According to Table 3 and Figure 4a, participants started significantly slower with OPTI (23.6 ± 15.5 WPM) compared to QWERTY (36.3 ± 10.8 WPM). Yet, starting from the 3rd repetition, they

were quickly able to exceed their corresponding QWERTY's typing speed. Still, OPTI only exhibited significantly better performance over QWERTY for repetitions 6 through 10. The curve for OPTI also starts to show a flattening trend from the 3rd repetition onwards. As for QWERTY, except for the 2nd repetition, the data seems to be more or less flat throughout.

The fastest average typing speed of 42.4 $\pm$ 13.4 WPM achieved using the OPTI layout was on the very last repetition. Yet, for QWERTY the fastest speed (38.6 $\pm$ 10.3 WPM) was already observed in the 2nd repetition. The fastest individual participant reached an average speed of 56.0 $\pm$ 11.6 WPM in the 10th repetition with OPTI. Similarly, for QWERTY, the fastest individual participant typing speed of 45.7 $\pm$ 9.90 WPM was observed in the 3rd repetition.

When comparing our results with Equations 1 and 2 from MacKenzie and Zhang's research [8], the average typing speed of the final repetition corresponds to 12.1 (20-22 minute long) sessions of normal/traditional training (i.e., training with different phrases) for QWERTY and 16.4 sessions for OPTI (see Table 3). Specifically, for QWERTY, where participants achieved their highest speed in the 2nd repetition, we estimate that reaching a speed of 38.6 WPM would require 15 sessions.

Interestingly, OPTI's typing speed performance showed a downward trend with increasing word length, as can be seen in Figure 5. Yet, the typing speed for QWERTY was more similar for all the word lengths (except for word length 12). For words that had a length of 5 or less, OPTI exhibited superior typing speed over QWERTY within the ten repetitions. For words with a word length of greater than 5, the curves crossed over each other, with QWERTY exhibiting superiority over OPTI.

A two-way RM ANOVA across keyboards and repetitions did not reveal significant differences for KSPC for the two dependent variables nor an interaction. As for MSD ER (see Table 2), with the exceptions of the 2nd and the 9th repetitions where OPTI exhibited significantly more errors than QWERTY (see Figure 4e), no other significant differences were found between the two keyboard layouts in the post-hoc analysis. Still, the MSD ER was throughout well below 2% for both the keyboards, allowing us to conclude that both KSPC (see Figure 4c) and MSD ER did not reveal any noteworthy trends.

Table 4: Phrase-level RM ANOVA results for the two Keyboard Layouts and ten Repetitions.

	Keyboards	Repetitions	Keyboards × Repetitions
WPM	$\boldsymbol{F_{1,13} = 8.45, p < 0.05,}$ $\boldsymbol{\eta^2 = 0.394}$	$\boldsymbol{F_{5.30,68.9} = 40.7, p < 0.001,}$ $\boldsymbol{\eta^2 = 0.758}$	$\boldsymbol{F_{9,117} = 29.0, p < 0.001,}$ $\boldsymbol{\eta^2 = 0.690}$
MSD ER	$\boldsymbol{F_{1,13} = 9.74, p < 0.01,}$ $\boldsymbol{\eta^2 = 0.428}$	$\boldsymbol{F_{9,117} = 2.58, p < 0.01,}$ $\boldsymbol{\eta^2 = 0.165}$	$F_{9,117} = 1.84, n.s.,$ $\eta^2 = 0.124$

4.5.2. ***Phrase-level Analysis*** *of WPM, KSPC, MSD ER, and Required Number of Traditional Training Sessions*

Tables 4 and 5 and Figure 4 present the results for the phrase-level performance metrics. According to Table 5 and Figure 4b, participants' typing speed was significantly slower in the first four repetitions with OPTI compared to QWERTY. Starting from the 5$^{\text{th}}$ repetition until the end, no significant differences in typing speed between OPTI and QWERTY were found. In other words, the two keyboard layouts exhibited similar performance from the 5$^{\text{th}}$ repetition onwards. In short, according to this measure **OPTI was never able to outperform QWERTY**.

The curve for OPTI showed a flattening trend starting from the 5$^{\text{th}}$ repetition. Although there was a slight increase in typing speed on the 9$^{\text{th}}$ repetition, the slope of the curve flattened out again on the 10$^{\text{th}}$ repetition. Conversely, the data for QWERTY was pretty flat throughout the ten repetitions.

The highest average typing speed of 35.2 ± 8.69 WPM attained using the OPTI layout occurred in the final repetition. For QWERTY, the highest speed of 35.2 ± 5.60 WPM was observed on the 8$^{\text{th}}$ repetition. Among individual participants, the fastest average speed reached 51.0 ± 5.08 WPM in the 10$^{\text{th}}$ repetition with OPTI. Similarly, with QWERTY, the highest individual participant typing speed (43.4 ± 5.43 WPM) was recorded on the 7$^{\text{th}}$ repetition.

When comparing our results with Equations 1 and 2 [8], the fastest average typing speed corresponds to 7.1 (20-22 minute long) sessions of traditional training for QWERTY and 9.2 sessions for OPTI (see Table 5).

Similar to our word-level KSPC results, a two-way RM ANOVA (again) did not reveal significant differences for phrase-level KSPC. As for MSD ER (see Table 4), OPTI exhibited significantly more errors than QWERTY in the first four and the 9$^{\text{th}}$ repetition (see Figure 4f). For repetitions 5-8 and 10, we did not observe significant differences between the two keyboard layouts.

Again, in all the repetitions for both keyboards, the MSD ER never reached 2%. This indicates that the participants were quite careful when typing on each of the two keyboards [40].

Table 5: Phrase-level WPM results for each repetition, overall totaling about 20 minutes of typing on each keyboard. Also, projections via Equations 1 and 2 of how many (22 minutes long) sessions or how much time would be required to reach the corresponding WPM with the traditional approach of typing different phrases during training.

	Typing Speed (WPM)		**Projected No. of Training Sessions**		**Projected Training Time (hours)**	
Repetition	**OPTI**	**QWERTY**	**OPTI**	**QWERTY**	**OPTI**	**QWERTY**
1	15.4 ± 4.67	31.8 ± 5.63	0.7	3.1	0.26	1.14
2	26.1 ± 8.99	33.9 ± 6.32	3.6	5.3	1.33	1.93
3	29.5 ± 10.3	34.0 ± 6.93	5.3	5.4	1.95	1.96
4	29.8 ± 9.34	34.3 ± 6.09	5.5	5.8	2.01	2.12
5	32.3 ± 9.81	33.8 ± 5.34	7.1	5.2	2.59	1.89
6	33.1 ± 9.83	34.7 ± 6.24	7.6	6.4	2.79	2.35
7	33.4 ± 10.4	34.8 ± 5.59	7.8	6.5	2.85	2.39
8	33.0 ± 9.67	35.2 ± 5.60	7.5	7.1	2.75	2.61
9	34.9 ± 8.46	34.2 ± 5.45	9.0	5.6	3.28	2.06
10	35.2 ± 8.69	34.8 ± 4.82	9.2	6.5	3.38	2.37

4.5.3. Learning Curves

We also fit a regression based on the power law of learning through the word-level and phrase-level WPM data [8, 18, 36], see Figures 6a and 6b. This yielded Equations 5 and 6 for word-level analysis:

$$WPM_{QWERTY} = 37.177 \times repetition^{0.0084}, R^2 = 0.0767 \quad (5)$$

$$WPM_{OPTI} = 24.688 \times repetition^{0.2533}, R^2 = 0.6358 \quad (6)$$

and Equations 7 and 8 for phrase-level analysis:

$$WPM_{QWERTY} = 31.682 \times repetition^{0.0424}, R^2 = 0.6762 \quad (7)$$

$$WPM_{OPTI} = 14.813 \times repetition^{0.3951}, R^2 = 0.7998 \quad (8)$$

In the above equations, *repetition* is the number of times the word is repeatedly typed, and R^2 represents the squared correlation coefficient.

4.5.4. Subjective Measures

11 out of the 14 participants preferred the QWERTY layout, and only 3 preferred OPTI. Justifications for selecting QWERTY included "*muscle memory*," "*habituation with [the] existing method*," and "*I use this layout every day, so I am pretty familiar with it*." Subjects who preferred OPTI mentioned "*the spacebars and the alphabet location, in particular, helped me [type] faster*," and "*there were four spacebars ... and the letters were better arranged*."

When participants were asked about their perceived ease of interaction with each keyboard using a 7-point Likert scale (1: very difficult, 7: very easy), the subjective feedback indicated a notable preference for QWERTY, which was perceived as significantly easier than OPTI ($t_{13} = 4.09, p < 0.001, d = 1.09$). Regarding frustration levels (1: very frustrating, 7: very satisfied), participants found QWERTY to be significantly less frustrating than OPTI. ($t_{13} = 2.73, p < 0.05, d = 0.73$). Similarly, in terms of mental fatigue (1: very fatiguing, 7: very relaxing; $t_{13} = 2.79, p < 0.05, d = 0.75$) and perceived speed (1: very slow, 7: very fast; $t_{13} = 3.40, p < 0.01, d = 0.91$), QWERTY again outperformed OPTI. The remaining two subjective measures, namely physical fatigue and perceived precision (1: very imprecise, 7: very precise), did not reveal significant differences. A graphical representation of the subjective outcomes is depicted in Figure 7.

4.6. Study 2 Discussion

In this study, we investigated whether repeatedly typing the same word provides an accurate estimation of trained user performance for unfamiliar keyboard layouts. To do this, we replicated MacKenzie and Zhang's study [8] with repeated word typing and analyzed the data following two different approaches, i.e., at the word- or phrase-level.

Results revealed that word-level analysis matches the findings of MacKenzie and Zhang [8] more closely. As previously mentioned, MacKenzie and Zhang [8] found participants were able to reach 44.3 WPM and about 40 WPM for OPTI and QWERTY, respectively, after 20 (20-22 minute) sessions of traditional training. The highest speed achieved in our study was 42.4 WPM for OPTI and 38.6 WPM for QWERTY (see Table 3 and Figure 4a). To reach such speeds, we can estimate (using Equations 1 and 2 [8]) that about 16 and 15 sessions, i.e., about 6 and 5.5 hours, of traditional training would be required, respectively (see Table 3). Moreover, and similar

to MacKenzie and Zhang's findings [8], our results also indicated that OPTI is superior to QWERTY.

Yet, we caution that word-level analysis is limited in terms of validity and reliability. The reason is that it does not take the word length distribution of a language into consideration and considers every word, whether of length two or ten, to have the same weight in the calculation of the averages. In other words, word-level analysis does not resemble the results of real-world keyboard usage well enough. Instead, a more valid and more widely comparable approach is a phrase-level analysis where the words are combined into their respective phrases.

However, the phrase-level analysis was unable to identify a superior keyboard nor did it provide accurate performance estimates (see Table 5 and Figure 4b). The main reason for this seems to be OPTI's deteriorating performance for longer words, as evident in Figure 5. Thus, taking the word length distribution in the phrase-level analysis into account resulted in different estimates. Our outcome also indicates that using only shorter words in a typing study [28] can bias the outcome of a study substantially (the OPTI data in Figure 5 is on average never below QWERTY for words up to 5 letters), and thus we do not recommend using only shorter words. Although not reported here, we also calculated word-level weighted averages of the typing speed, where the weights were set to the lengths of each word. From this, we found quite similar results to that of our phrase-level analysis presented in Figure 4b, matching the pattern of the curves similarly well.

The learning curves for word-level and phrase-level analysis in Equations 5, 6, 7, and 8 also show that phrase-level analysis fits the performance data better than word-level analysis, as evident through the higher R^2 values. This again shows that phrase-level analysis is potentially a more valid approach than word-level analysis.

One reason the curve for QWERTY was fairly flat in both Studies 1 and 2 (see Figures 3 and 6) is attributable to the fact that the QWERTY layout did not involve much learning, due to participants' prior familiarity with the layout. Thus, QWERTY was also highly preferred and rated as evident in the subjective outcomes of this study.

5. Guided Evaluation Method (GEM)

The findings of Study 2 discourage the use of the repeated word typing approach for text entry studies, as this approach was unable to accurately

estimate trained user performance (if word length distributions were correctly accounted for). Thus, we introduce here a new approach: the **Guided Evaluation Method (GEM)**. The GEM is designed to be less resource-consuming than traditional evaluation approaches. To reduce a novice's visual search time for finding keys on an unfamiliar keyboard [33, 34], the GEM builds on the idea of providing visual clues [41, 42] and *guides* the user by highlighting all the characters of the target **word**, more specifically, the keys that are yet to be typed (see Figure 1). Thus, with a target word of "ELECTIONS" and as shown in Figure 1a, when "EL" has already been typed, the letter 'L' is no longer highlighted as it does not exist elsewhere in the remaining part of the target word. If participants make a mistake, keys are not de-highlighted from the mistake onwards. This helps participants realize that a mistake had occurred earlier in the word and does not require them to constantly check whether the typed and the target text matched.

Similar to Study 2, once the target word was typed, we asked participants to type in a space character to denote the end of the word. Also, we (again) randomly chose complete phrases from MacKenzie and Soukoreff's set [61], but (again) presented only one individual word at a time.

We chose to present only a single word at a time because we did not want to overburden our novice participants with the need to remember the whole phrase – after all, they are simultaneously dealing with the unfamiliar OPTI keyboard layout, which already increases cognitive overhead. Thus, having to memorize only a single word (instead of a phrase) is more likely to reveal representative typing speeds, as participants need to memorize less text at a time. Furthermore, typing only a single word at a time also requires less going back and forth to check whether the typed and the target text match, again avoiding extra time that is not representative of skilled typing.

In our first pilot study during the development process of the GEM, we tried to highlight only the next letter, i.e., highlight a single key/letter of the target word at a time. Yet, this did not work well, as our novice participants took a lot of time to first perceive and react to each new highlight, i.e., the time needed to process a new visual stimulus [74, 75], before pressing the corresponding key. As the time interval required to respond to a new visual stimulus is not representative of trained/expert user behavior, we decided to use a different approach.

In a subsequent pilot study, we asked participants to first *plan* the typing pattern of each target word before they started typing that word without highlighting any letters. We encouraged them to take as long as they needed

to plan the typing pattern. Once the planning was done, they were asked to type the target word as quickly as possible. Yet, this approach also did not work well, as we noticed that participants were not able to remember the typing pattern beyond four or five letters of the target word. In other words, participants frequently got "lost" after they typed four/five letters and started acting like pure novices, i.e., scanned the entire keyboard layout for each next target letter.

To address this issue and to simultaneously compensate for the above-mentioned per-key reaction times, we asked participants to plan the typing pattern for each word before typing it and also highlight all the keys of the target word on the keyboard that are still to be typed. Highlighting all keys to be typed permits us to de-highlight all keys that have already been typed — as long as they do not exist elsewhere in the target word. This significantly reduces the visual search space and thus (largely) mitigates the issue of participants getting lost after typing four or five letters.

In summary, GEM comprises three main features — typing only a single word at a time, planning before typing, and reducing the search space by highlighting all necessary keys but also de-highlighting keys that have already been typed. These three properties of the GEM allow us to substantially reduce learning effects and thus the need for a longitudinal study, which enables us to use the GEM to directly compare different keyboard layouts, e.g., OPTI vs. QWERTY. In other words, GEM-based user studies with *novice users* have the potential to enable designers to *quickly* estimate a layout's performance with trained users.

6. User Study 3 – GEM

In this study, we again replicated MacKenzie and Zhang's study [8], but this time with the proposed GEM approach. This choice, just as in Studies 1 and 2, allowed us to compare our results with their findings [8] and, in the process, helped us to validate the GEM approach. Here, we describe the details of the study and how we evaluated the two keyboard layouts, OPTI and QWERTY, with the GEM.

6.1. Participants

14 participants (10 male, 4 female), aged 30.3 ± 5.01 years, took part in this study. All of them had experience with typing on a physical and/or soft QWERTY keyboard for over nine years, except one who has been typing for

3-5 years. For all participants, this study was the first time they typed on an OPTI keyboard. Each participant was compensated with $10 for their time.

6.2. Apparatus, Keyboard Designs, and Procedure

The same apparatus and keyboard design as in Studies 1 and 2 were used, except for the addition of the key highlighting associated with the GEM (see Figure 1). At the beginning of the study, participants filled out a consent form and a demographic questionnaire. Participants then typed eleven phrases on each of the two keyboards. We again instructed participants to memorize each target word (and if needed its spelling) and to plan the typing pattern of each word before they started typing it. We encouraged them to take as long as they needed during the planning phase. We (again) asked participants to type each word of the randomly chosen phrases [61] in sequence, by revealing only a single word at a time.

Participants were not given practice trials. Instead, we again considered the first phrase as practice and discarded it from the analysis [14, 67]. Similar to Study 2, at the end of the experiment, participants filled out another short questionnaire. On average, participants typed for about 5 minutes on each of the two keyboards. In total, the experiment took less than 30 minutes, including the demographic pre-questionnaire and the post-questionnaires. Everything else was the same as in the previous two studies.

6.3. Performance Metrics

Other than the **phrase-level** WPM, KSPC, and MSD ER measures used in Studies 1 and 2, we also recorded the planning time in this study to evaluate the text entry performance of the two keyboard layouts. The *Planning Time* is the time taken by the participants to plan the typing pattern for a word in the GEM. In other words, it is the time taken from the presentation of a word to the first key press for entering said word. Note that the planning time is not taken into account in the calculation of WPM. Similar to Studies 1 and 2, the WPM measure just considers the time taken by the participants to type the word — from the first key press to the last key press (i.e., the spacebar).

6.4. Experimental Design

We used a within-subjects design in this study. As dependent variables, we measured participants' WPM, KSPC, MSD ER, and planning time. For each of these performance metrics, we calculated the overall average of the

ten typed phrases, i.e., ignoring the first practice phrase. As the overall performance data can be confounded by the properties of each individual phrase (see below), we used a 5-phrase moving average (5-PMAvg) for WPM, KSPC, MSD ER, and planning time to analyze participants' performance over time. Thus, for the overall averages and the subjective measures, we had a single independent variable – the keyboard layout. As for the 5-PMAvg, we had two independent variables – the keyboard layout and the phrase number.

The reason we chose a moving average over a simple average for each individual phrase is that the latter suffers a similar problem as the repeatedly typing the same phrase approach [32] – it does not represent real keyboard usage as the average is calculated based on typing just a single phrase. Thus, confounding factors like the characters, length, and number of words of the particular phrase could have a higher impact on the results than desirable. To mitigate this issue, we used a moving average for each of the four performance metrics. For the first four phrases, we calculated the moving average only up to that point. For example, the moving average for the third phrase is thus the average of only the first three phrases. We also chose to calculate the moving averages over *five* phrases because several existing studies, e.g., [11, 25, 24], define a single session to comprise of typing five phrases, and we wanted to support comparability of our results with the literature.

6.5. Results

We analyzed the data using t-tests for the overall averages and the subjective measures. For the 5-PMAvg, we analyzed the data using two-way RM ANOVA. For both the t-tests and RM ANOVA, we analyzed the data in the same manner as in Study 2. We used Bonferroni corrections for the ANOVAs to prevent the inflation of type I errors.

6.5.1. WPM, KSPC, MSD ER, and Required Number of non-GEM Sessions

According to Figure 8a, participants' overall average typing speed across all ten phrases was significantly faster ($t_{13} = 4.15, p < 0.001, d = 1.11$) for OPTI ($44.6 \pm 2.60$ WPM) compared to QWERTY (39.3 ± 1.61 WPM). The fastest participant was able to type at an overall average of 58.7 ± 10.3 WPM with OPTI and 46.5 ± 7.84 WPM with QWERTY. As per Figures 8b and 8c, no significant differences were observed for the overall KSPC (OPTI: 1.020 $\pm$ 0.013, QWERTY: 1.034 $\pm$ 0.022) and MSD ER (OPTI: 0.29 $\pm$ 0.39%, QWERTY: 0.13 $\pm$ 0.27%) measures.

When looking at how typing speed varies over time (see Tables 6 and 7), the results of the 5-PMAvg reveal that participants started typing significantly faster with OPTI starting from the 6^{th} phrase (see Figure 9a). The fastest participant was able to type at a 5-PMAvg of 64.8 ± 7.50 WPM with OPTI and 48.9 ± 7.50 WPM with QWERTY. No significant differences for a 5-PMAvg of KSPC and MSD ER were observed (see Figures 9b and 9c).

When we compare our overall average WPM results for the GEM with extrapolations through Equations 1 and 2 (which do not include the GEM), it would take 19.1 (20-22 minute) sessions for OPTI (i.e., 7.02 hours of training) and 17.5 sessions for QWERTY (i.e., 6.42 hours of training) for participants to reach the same typing speed. For the 5-PMAvg WPM results, the corresponding number of MacKenzie and Zhang's sessions and total training time [8], is presented in Table 7.

Table 6: 5-phrases Moving Average RM ANOVA results for the two Keyboard Layouts and ten Phrases.

	Keyboards	Phrases	Keyboards × Phrases
WPM	$\boldsymbol{F_{1,9} = 7.71, p < 0.05,}$ $\boldsymbol{\eta^2 = 0.461}$	$F_{2.18,19.6} = 0.63, n.s.,$ $\eta^2 = 0.066$	$\boldsymbol{F_{3.61,32.5} = 3.04, p < 0.01,}$ $\boldsymbol{\eta^2 = 0.253}$
Planning Time	$\boldsymbol{F_{1,9} = 39.1, p < 0.001,}$ $\boldsymbol{\eta^2 = 0.813}$	$F_{2.91,26.2} = 1.40, n.s.,$ $\eta^2 = 0.135$	$F_{3.06,27.5} = 0.58, n.s.,$ $\eta^2 = 0.060$

6.5.2. Planning Time Analysis

The comparison of the overall average planning time for all the phrases showed that participants took significantly longer ($t_{13} = 4.61, p < 0.001, d = 1.23$) to plan their typing with the OPTI keyboard (21.4 ± 2.65 seconds) compared to QWERTY (12.9 ± 1.84 seconds; see Figure 8d). When looking at a 5-PMAvg of planning time (see Table 6), OPTI required significantly more time to plan for all phrases over QWERTY (see Figure 10a). Further, Figure 10b shows that the planning time increased with increasing word length.

6.5.3. Subjective Measures

Among our 14 participants, 10 preferred QWERTY, 3 preferred OPTI, and 1 mentioned that both were the same. Example reasons for the choice of QWERTY were "*Using [QWERTY] for more than 10 years*", "*Familiarity, wide availability*", it is the "*same as a laptop/phone keyboard*", and it "*felt natural and easy.*" Subjects who preferred OPTI mentioned "*the keyboard*

Table 7: 5-phrases Moving Average of WPM results for each phrase, overall totaling about 5 minutes of typing on each keyboard. Also, projections via Equations 1 and 2 of how many (22 minutes long) sessions or how much time would be required to reach the corresponding WPM with the traditional approach of typing different phrases during training.

	Typing Speed (WPM)		**Projected No. of Training Sessions**		**Projected Training Time (hours)**	
Phrase	**OPTI**	**QWERTY**	**OPTI**	**QWERTY**	**OPTI**	**QWERTY**
1	40.5 ± 7.21	41.0 ± 8.59	14.2	24.7	5.20	9.07
2	41.7 ± 9.18	41.2 ± 7.47	15.5	25.6	5.69	9.39
3	41.9 ± 9.28	40.0 ± 7.50	15.7	20.0	5.77	7.33
4	41.9 ± 9.46	39.6 ± 6.58	15.8	18.4	5.79	6.75
5	41.9 ± 9.75	39.4 ± 5.76	15.8	17.7	5.81	6.51
6	44.0 ± 9.31	38.8 ± 5.41	18.4	15.7	6.73	5.74
7	44.2 ± 7.61	38.3 ± 4.29	18.6	14.1	6.84	5.18
8	45.8 ± 7.77	38.8 ± 3.75	20.8	15.8	7.63	5.79
9	45.3 ± 7.38	39.1 ± 4.48	20.0	16.6	7.35	6.09
10	45.6 ± 5.86	39.0 ± 4.34	20.5	16.4	7.50	6.01

responds faster", "*Letters are closer*", and "*Four [spacebars] ... helped me to get to the next word faster.*" The participant who did not have a particular preference said that "*both are easy to learn.*"

When asked about the ease of interaction with each keyboard on a 7-point Likert scale (1: very difficult, 7: very easy), the subjective responses showed that QWERTY was perceived to be easier than OPTI ($t_{13} = 2.26, p < 0.05, d = 0.61$). In terms of frustration (1: very frustrating, 7: very satisfied), QWERTY was less frustrating than OPTI ($t_{13} = 2.23, p < 0.05, d = 0.60$). Similarly, for mental fatigue (1: very fatiguing, 7: very relaxing), QWERTY was again better than OPTI ($t_{13} = 3.31, p < 0.01, d = 0.88$). The other three subjective measures, i.e., physical fatigue, perceived speed (1: very slow, 7: very fast), and precision (1: very imprecise, 7: very precise) did not exhibit significant differences. A plot of the subjective results is presented in Figure 10c.

6.6. Study 3 Discussion

In this study, we investigated whether we can accurately estimate trained user-level performance for unfamiliar keyboards with the GEM approach.

To evaluate the GEM, we first compare our results with those of MacKenzie and Zhang's study [8]. Our findings with the GEM are quite close: 44.6

WPM with GEM vs. 44.3 WPM in MacKenzie and Zhang's 20th session for OPTI, respectively 39.3 vs. 40 WPM for QWERTY. These results demonstrate that *a ≈5-minute session with the GEM for each keyboard is able to identify very similar results to those that were reached after about 7 hours of a longitudinal study per participant without the GEM.* We also confirmed that OPTI can reach speeds that are overall significantly faster than QWERTY (see Figure 8a). We cannot directly compare this finding with MacKenzie and Zhang [8], as the authors did not report such statistical differences in their work.

In terms of the 5-PMAvg, OPTI exhibited significant differences compared to QWERTY starting from the 6th phrase (see Figure 9a and Table 7). Consequently, we recommend running the GEM for at least *six* phrases but acknowledge that this number can vary depending on the exact interaction/pointing/selection technique. Thus, the number of phrases used with the GEM is best determined by running pilots as more phrases may be needed, e.g., to differentiate between very similar layouts/selection techniques.

The highest 5-PMAvg typing speed for OPTI was 45.8 WPM on phrase 8, while it was 41.2 WPM on phrase 2 for QWERTY. Thus, in the context of touch-based typing and *in comparison to a longitudinal study, our results validate that the GEM is a lot more efficient for estimating the trained user performance of a text entry system.* However, the results with the GEM did not get close to the expert-level performance of 60.7 WPM with OPTI and 44.8 WPM with QWERTY *predicted* by MacKenzie and Zhang [8] nor to the text entry rates of 55.6-58.93 WPM reported for experts in other literature [76, 77]. Thus, we can only state that the GEM estimates the performance of trained users, i.e., users trained via a typical text entry longitudinal study, and *not* that of true experts.

It is also illustrative to compare the predictions from the GEM with the model proposed by MacKenzie and Zhang's [8]. To reach the overall average typing speed achieved with the GEM, participants would need to train for **19.1 and 17.5 non-GEM sessions** (each 20-22 minutes), i.e., train for **7.02 and 6.42 hours**, for OPTI and QWERTY with MacKenzie and Zhang's [8] approach, respectively.

Similarly, for the highest 5-PMAvg result, participants would need to go through **20.8 non-GEM sessions (i.e., 7.63 hours)** with OPTI and **17.7 sessions (i.e., 6.51 hours)** with QWERTY (see Table 7). These results also illustrate *how much time the GEM can save for empirically estimating*

trained user-level performance on novel/unfamiliar keyboard layouts.

As shown in Figure 9a and Table 7, the typing speed of participants with the OPTI keyboard did improve over time. However, the speed decreased with increasing time for QWERTY. This is probably explained by the slightly higher KSPC for QWERTY compared to OPTI (see Figure 9b). Based on our observations, the potential reason for the higher KSPC, although not significantly different from OPTI, is that participants felt more confident with this keyboard layout. Thus, they tended to type faster, which led to more mistakes [28] and, therefore, more corrections [16], which slowed their overall performance [46]. Still, the typing speed for QWERTY was consistently above 38 WPM, see Table 7, unlike our word-level analysis in Study 2 (see Figure 4a and Table 3). In other words, the performance observed with the GEM is quite close to the previously reported speed of about 40 WPM for regular QWERTY users [8].

The GEM also did not counteract participants' years of experience with the QWERTY layout. This is evident in the results of the subjective measures (see Figure 10c) for each keyboard, where most of the participants preferred QWERTY over OPTI. QWERTY was also rated significantly better than OPTI for ease of interaction, frustration, and mental fatigue.

Beyond participants' familiarity with the QWERTY layout, another reason why QWERTY was rated better is likely the increased effort (and time) required to plan the typing pattern for the OPTI keyboard. Not only was the overall planning time significantly less for QWERTY (see Figure 8d), it also took participants significantly less time to plan throughout the ten phrases compared to OPTI (see Figure 10a). We believe that this higher planning time for OPTI is one of the key reasons that led to participants exhibiting a faster typing speed within the GEM relative to QWERTY. After all, trained/expert users will typically either explicitly plan execution in advance or implicitly use muscle memory, e.g., [78]. Thus, we believe this does not invalidate the GEM as an evaluation method — in contrast, it emphasizes that the GEM uses a methodology that is more commensurate with trained user behaviors. Moreover, according to Figure 10b, the time required to plan to type a word on the OPTI keyboard also seems to increase as the length of the word to be typed increases, and notably so for words that are 9 letters or longer. Similar observations were made for typing speed in our study, where the WPM decreased with increasing word length for both OPTI and QWERTY.

7. General Discussion and Future Work

In this paper, we first showed that the approach of repeatedly typing the same phrase could estimate trained user-level text entry performance with data gathered for more than three days of usage from each participant. Yet, this is still effectively a longitudinal study. We then investigated the repeated word typing approach. Results indicated that when the data were analyzed at the phrase-level, and not at the word-level (which ignores word length distributions), this approach failed to estimate trained user performance accurately. Towards more efficient evaluations of novel keyboard layouts, we then proposed and validated a novel approach, the Guided Evaluation Method (GEM), which empirically estimates how fast trained users can type with a new layout through a study with novice users in a matter of minutes. Thus, the GEM greatly reduces the need for longitudinal studies in text entry research.

In Study 2, all significant differences of the main objective measures exhibited large effect sizes ($\eta^2 > 0.14$; see Tables 2 and 4). Similarly, in Study 3, we found significant differences between the two keyboards for overall average typing speed with a large effect size ($d = 1.11$) with 14 participants as well, a number consistent with much other work [79]. In terms of 5-PMAvg of typing speed, the effect sizes were also large ($\eta^2 > 0.14$; see Table 6). These robust findings make us believe that our results are likely replicable.

In Study 2, the phrase-level model for OPTI, i.e., Equation 8, achieved the highest R^2 value of 0.7998. This equation predicts that it would take approximately 16 repetitions of typing the same word for participants to reach speeds of 44.3 WPM, as reported previously [8]. However, given the flattening trend of the OPTI curve starting from the 5^{th} repetition, we have strong doubts about whether the prediction would actually play out in reality. Based on our observations of the participants and listening to the auditory feedback provided by the system on each key press, participants tended to get into a rhythm by the $5^{\text{th}}/6^{\text{th}}$ repetition. For example, participants typed the first 3-5 letters in one go, paused ever so slightly, and typed the rest of the letters in another go. This rhythm was continued for the remaining repetitions. To mitigate this, we recommend future work to introduce a planning phase every five repetitions, similar to our GEM approach. We speculate that planning the typing pattern after every fifth repetition would improve participants' performance further, possibly even matching our learning models in Equations 7 and 8. More importantly, this could even provide more

accurate phrase-level estimates of trained user performance.

Although the GEM was quite accurate in estimating trained user-level text entry performance, it was unable to predict the performance of true experts, i.e., users who have years-long training. Still, the GEM is versatile enough to be easily integrated into different types of prototype text entry systems that utilize different interaction techniques, e.g., touch/tap, swipe, eye-gaze-based techniques, and others. Thus, researchers can now use the GEM to focus directly on the training of an interaction technique over multiple sessions and can (largely) ignore the potential confound of layout learning. A longitudinal study with the GEM could then potentially answer the question of whether the true expert typing speeds predicted by previous work [5, 8] for different layouts could actually be achieved, something that is very difficult to do with other techniques as participants otherwise need to spend (even more) time to learn both the layout and the interaction technique. Thus, we recommend future work to investigate if and how quickly users can achieve the speed of a true expert with the GEM.

As mentioned before, the GEM relies on three main features — typing a single word at a time, planning before typing, and reducing the search space by highlighting all necessary keys but also de-highlighting keys that have already been typed. Although we found in our pilots that each of these individual features is not sufficient to predict the performance of a trained user on its own, it would still be interesting to investigate in the future how much each of these three components contributes to the successful predictions of the GEM approach.

Through other unreported pilot studies that we performed before the studies reported here, we learned quickly that the GEM is *not* a good means for inducing *learning* of a keyboard layout, specifically because the GEM – as presented above – involves only typing a few phrases. We also need to acknowledge here that the GEM involves a typing experience that is not the same as real-life typing. Thus, we highlight that the GEM approach is currently only designed to be an evaluation method to quickly estimate trained user performance and/or compare different keyboard layouts and that *the GEM in its current form should not be used as a training method to learn a new layout.* Still, it might be interesting to analyze in future work if something like the GEM, or more specifically highlighting target keys, could have benefits for training users on unfamiliar keyboard layouts.

To enable direct comparisons with existing studies, we chose to present phrases word by word with the GEM. Still, further work should investigate

the influence of word length on planning time and typing speed on unfamiliar keyboards. Additionally, a combination of repeatedly typing the same word/phrase and the GEM might allow for an even better/faster trained/expert user performance estimation. Again, further studies are required to verify this.

8. Limitations

Although the GEM provides an accurate estimation of the text entry speed for a given layout, one limitation of the approach is its bias towards a lower error rate. Previous work has demonstrated that expert users tend to make more mistakes when typing very fast [28] and thus require more corrective actions [16]. Yet, according to our data shown in Figures 8c and 9c, our participants made substantially fewer errors compared to, for example, the trained users of MacKenzie and Zhang [8] or the corrective actions performed by the trained users of Majaranta et al. [16]. We believe that the reason why users made fewer mistakes with the GEM is that, as the keys are highlighted in GEM and unlike real-world typing, users are less likely to make mistakes that involve keys outside that limited sets of highlighted keys. Therefore, we recommend using GEM predominantly to make rapid *a priori* predictions for the text entry speed potential of a text entry technique in the early stages of the development process.

Another approach to estimating expert user performance for text entry is to use model-based approaches, such as the KLM model [80] or variants [81], including those adapted to touch screens [82]. Yet, such models are targeted (only) at expert performance, i.e., focus mostly on motor movement time, and do not take learning into account nor model the performance of users that have received only a specified amount of training, i.e., trained users. On the surface, it seems feasible to adopt these approaches to take learning into account, e.g., by modeling the decrease of the mental effort (and thus time) to recall a character's location on the keyboard. Yet, this is a simplistic approach, as it assumes that recall is always perfect. Previous models for text entry learning built on cognitive architectures for human memory, such as ACT-R [83], e.g., modeling memorization and also the interference effect from distractors encountered during learning [84, 85, 86, 87].

While the repeated word/phrase typing approach involves a memory recall component, it relies predominantly on the immediate preceding trials. In contrast, the GEM bypasses the issue of letter location recall by mov-

ing memory recall and/or visual scanning (mostly) into the planning phase before the word is entered. In essence, this a-priori planning improves our novice participants' execution time, yielding results that are close to trained user performance. Still, we (again) acknowledge that these performance estimates are achieved by the GEM by simulating a typing experience that is not the same as real-life typing. We further recognize that the GEM involves a mechanism that is somewhat different from the behavior of a real trained user, e.g., because trained users typically remember letter locations while the GEM only highlights them, and thus, users have to remember them only for the current word.

Another limitation of this work is that to validate both the repeated word/phrase typing approaches as well as the GEM, we compared our results to that of MacKenzie and Zhang [8]. Although we tried to replicate their work [8] as closely as possible, we still had to make some modifications that potentially confounded the results to some extent. For example, the keyboard design was slightly modified as the dimension of the keys had to be reduced a bit (0.2 cm), and a small gap (0.1 cm) between the keys needed to be introduced. Another difference was we asked participants to use their dominant hand's index finger instead of the stylus used by MacKenzie and Zhang [8]. The difference in technological advances since their work [8] and a smaller user sample, though not uncommon in longitudinal studies [20, 35], could also be factors as well. Still, we believe that these factors do not invalidate our work. Not only were our findings very similar to previous work [5, 8, 38], but also the fact that our absolute novice participants were able to type significantly faster with OPTI compared to QWERTY starting from just the 6^{th} phrase makes us believe that our study is very comparable. This is evidence that the GEM unquestionably mitigates participants' lack of layout knowledge. Still, we encourage future work to conduct a longitudinal study with the same apparatus and participants to further validate our work.

9. Conclusion

To reduce the logistical bottleneck of longitudinal studies in text entry research for reliably estimating the performance potential of a given text entry system, we investigated multiple approaches here. First, we showed that the popular approach of repeatedly typing the same phrase would still require training participants for more than three days to estimate trained user text entry performance. Therefore, it does not eliminate the need for a longitudi-

nal study. Next, we found that the approach of repeatedly typing the same word does not provide accurate estimates of trained user performance when the data is analyzed at a phrase-level, which is more valid than analysis at the word-level, where word length distributions are ignored.

We then proposed a novel approach, the Guided Evaluation Method (GEM), and showed that it can accurately estimate the performance of trained users in a user study with novice users in just a matter of minutes. Thus, the GEM greatly reduces the need for longitudinal studies and enables researchers to quickly perform text entry studies with different interaction techniques and layouts. Given this, we conclude by stating that the GEM has the potential to be a game changer in the investigation of novel keyboard layouts and interaction techniques and maybe even text entry research in general.

References

[1] P. Majaranta, Communication and text entry by gaze, in: Gaze interaction and applications of eye tracking: Advances in assistive technologies, IGI Global, 2012, pp. 63–77. doi:10.4018/978-1-61350-098-9.ch008.

[2] C. Kumar, R. Menges, D. Müller, S. Staab, Chromium based framework to include gaze interaction in web browser, in: Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2017, p. 219–223. doi:10.1145/3041021.3054730.
URL `https://doi.org/10.1145/3041021.3054730`

[3] C. Pandarinath, P. Nuyujukian, C. H. Blabe, B. L. Sorice, J. Saab, F. R. Willett, L. R. Hochberg, K. V. Shenoy, J. M. Henderson, High performance communication by people with paralysis using an intracortical brain-computer interface, eLife 6 (2017) e18554. doi:10.7554/eLife.18554.

[4] M. E. Mott, S. Williams, J. O. Wobbrock, M. R. Morris, Improving dwell-based gaze typing with dynamic, cascading dwell times, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17, Association for Computing Machinery, New York,

NY, USA, 2017, p. 2558–2570. doi:10.1145/3025453.3025517.
URL https://doi.org/10.1145/3025453.3025517

[5] J. Rick, Performance optimizations of virtual keyboards for stroke-based text entry on a touch-based tabletop, in: Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology, UIST '10, Association for Computing Machinery, New York, NY, USA, 2010, p. 77–86. doi:10.1145/1866029.1866043.
URL https://doi.org/10.1145/1866029.1866043

[6] T. J. Dube, A. S. Arif, Text entry in virtual reality: A comprehensive review of the literature, in: M. Kurosu (Ed.), Human-Computer Interaction. Recognition and Interaction Technologies, Springer International Publishing, Cham, 2019, pp. 419–437. doi:10.1007/978-3-030-22643-5$_3$3.

[7] M. Speicher, A. M. Feit, P. Ziegler, A. Krüger, Selection-based text entry in virtual reality, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 1–13. doi:10.1145/3173574.3174221.
URL https://doi.org/10.1145/3173574.3174221

[8] I. S. MacKenzie, S. X. Zhang, The design and evaluation of a high-performance soft keyboard, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '99, Association for Computing Machinery, New York, NY, USA, 1999, p. 25–31. doi:10.1145/302979.302983.
URL https://doi.org/10.1145/302979.302983

[9] P. O. Kristensson, K. Vertanen, The potential of dwell-free eye-typing for fast assistive gaze communication, in: Proceedings of the Symposium on Eye Tracking Research and Applications, ETRA '12, Association for Computing Machinery, New York, NY, USA, 2012, p. 241–244. doi:10.1145/2168556.2168605.
URL https://doi.org/10.1145/2168556.2168605

[10] X. Lu, D. Yu, H. N. Liang, W. Xu, Y. Chen, X. Li, K. Hasan, Exploration of hands-free text entry techniques for virtual reality, in: 2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), IEEE, 2020, pp. 344–349. doi:10.1109/ISMAR50242.2020.00061.

[11] C. H. Morimoto, J. A. T. Leyva, A. Diaz-Tula, Context switching eye typing using dynamic expanding targets, in: Proceedings of the Workshop on Communication by Gaze Interaction, COGAIN '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 1–9. doi:10.1145/3206343.3206347.
URL https://doi.org/10.1145/3206343.3206347

[12] A. Kurauchi, W. Feng, A. Joshi, C. H. Morimoto, M. Betke, Swipe&switch: Text entry using gaze paths and context switching, in: Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology, UIST '20 Adjunct, Association for Computing Machinery, New York, NY, USA, 2020, p. 84–86. doi:10.1145/3379350.3416193.
URL https://doi.org/10.1145/3379350.3416193

[13] W. Feng, J. Zou, A. Kurauchi, C. H. Morimoto, M. Betke, Hgaze typing: Head-gesture assisted gaze typing, in: ACM Symposium on Eye Tracking Research and Applications, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1–11.
URL https://doi.org/10.1145/3448017.3457379

[14] C. Yu, K. Sun, M. Zhong, X. Li, P. Zhao, Y. Shi, One-dimensional handwriting: Inputting letters and words on smart glasses, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 71–82. doi:10.1145/2858036.2858542.
URL https://doi.org/10.1145/2858036.2858542

[15] F. C. Y. Li, R. T. Guy, K. Yatani, K. N. Truong, The 1line keyboard: A qwerty layout in a single line, in: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, UIST '11, Association for Computing Machinery, New York, NY, USA, 2011, p. 461–470. doi:10.1145/2047196.2047257.
URL https://doi.org/10.1145/2047196.2047257

[16] P. Majaranta, U.-K. Ahola, O. Špakov, Fast gaze typing with an adjustable dwell time, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '09, Association for Computing Machinery, New York, NY, USA, 2009, p. 357–360. doi:10.1145/1518701.1518758.
URL https://doi.org/10.1145/1518701.1518758

[17] O. Tuisku, P. Majaranta, P. Isokoski, K.-J. Räihä, Now dasher! dash away! longitudinal study of fast text entry by eye gaze, in: Proceedings of the 2008 Symposium on Eye Tracking Research & Applications, ETRA '08, Association for Computing Machinery, New York, NY, USA, 2008, p. 19–26. doi:10.1145/1344471.1344476.
URL https://doi.org/10.1145/1344471.1344476

[18] J. P. P. Jokinen, S. Sarcar, A. Oulasvirta, C. Silpasuwanchai, Z. Wang, X. Ren, Modelling learning of new keyboard layouts, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 4203–4215. doi:10.1145/3025453.3025580.
URL https://doi.org/10.1145/3025453.3025580

[19] E. Clarkson, J. Clawson, K. Lyons, T. Starner, An empirical study of typing rates on mini-qwerty keyboards, in: CHI '05 Extended Abstracts on Human Factors in Computing Systems, CHI EA '05, Association for Computing Machinery, New York, NY, USA, 2005, p. 1288–1291. doi:10.1145/1056808.1056898.
URL https://doi.org/10.1145/1056808.1056898

[20] H. Hakoda, B. Shizuki, J. Tanaka, Qaz keyboard: Qwerty based portrait soft keyboard, in: A. Marcus (Ed.), Design, User Experience, and Usability: Technological Contexts, Springer International Publishing, Cham, 2016, pp. 24–35. doi:10.1007/978-3-319-40406-6_3.

[21] S. J. Castellucci, I. S. MacKenzie, M. Misra, L. Pandey, A. S. Arif, Tiltwriter: Design and evaluation of a no-touch tilt-based text entry method for handheld devices, in: Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia, MUM '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–8. doi:10.1145/3365610.3365629.
URL https://doi.org/10.1145/3365610.3365629

[22] K. Lyons, T. Starner, D. Plaisted, J. Fusia, A. Lyons, A. Drew, E. W. Looney, Twiddler typing: One-handed chording text entry for mobile phones, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '04, Association for Computing Machinery, New York, NY, USA, 2004, p. 671–678. doi:10.1145/985692.985777.
URL https://doi.org/10.1145/985692.985777

[23] D. Rough, K. Vertanen, P. O. Kristensson, An evaluation of dasher with a high-performance language model as a gaze communication method, in: Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces, AVI '14, Association for Computing Machinery, New York, NY, USA, 2014, p. 169–176. doi:10.1145/2598153.2598157.
URL https://doi.org/10.1145/2598153.2598157

[24] C. Kumar, R. Hedeshy, I. S. MacKenzie, S. Staab, Tagswipe: Touch assisted gaze swipe for text entry, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1–12. doi:10.1145/3313831.3376317.
URL https://doi.org/10.1145/3313831.3376317

[25] R. Hedeshy, C. Kumar, R. Menges, S. Staab, Hummer: Text entry by gaze and hum, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1–11.
URL https://doi.org/10.1145/3411764.3445501

[26] W. Xu, H.-N. Liang, Y. Zhao, T. Zhang, D. Yu, D. Monteiro, Ringtext: Dwell-free and hands-free text entry for mobile head-mounted displays using head motions, IEEE Transactions on Visualization and Computer Graphics 25 (5) (2019) 1991–2001. doi:10.1109/TVCG.2019.2898736.

[27] S. Ghosh, A. Joshi, M. Joshi, N. Emmadi, G. Dalvi, S. Ahire, S. Rangale, Shift+tap or tap+longpress? the upper bound of typing speed on inscript, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 2059–2063. doi:10.1145/3025453.3025944.
URL https://doi.org/10.1145/3025453.3025944

[28] X. Bi, S. Zhai, Ijqwerty: What difference does one key change make? gesture typing keyboard optimization bounded by one key position change from qwerty, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 49–58. doi:10.1145/2858036.2858421.
URL https://doi.org/10.1145/2858036.2858421

[29] K. Vertanen, H. Memmi, J. Emge, S. Reyal, P. O. Kristensson, Velocitap: Investigating fast mobile text entry using sentence-based decoding

of touchscreen keyboard input, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15, Association for Computing Machinery, New York, NY, USA, 2015, p. 659–668. doi:10.1145/2702123.2702135.
URL https://doi.org/10.1145/2702123.2702135

[30] K. A. Ericsson, R. T. Krampe, C. Tesch-Römer, The role of deliberate practice in the acquisition of expert performance., Psychological review 100 (3) (1993) 363.

[31] M. H. Urbina, A. Huckauf, Dwell time free eye typing approaches, in: Proceedings of the 3rd Conference on Communication by Gaze Interaction, COGAIN '07, Association for Computing Machinery, New York, NY, USA, 2007, pp. 65–70.
URL https://wiki.cogain.org/images/e/e5/COGAIN2007Proceedings.pdf

[32] A. K. Mutasim, M. Hudhud Mughrabi, A. U. Batmaz, W. Stuerzlinger, Does repeatedly typing the same phrase provide a good estimate of expert text entry performance?, in: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1–8. doi:https://doi.org/10.1145/3544549.3585647.

[33] I. S. MacKenzie, S. X. Zhang, An empirical investigation of the novice experience with soft keyboards, Behaviour & Information Technology 20 (6) (2001) 411–418. arXiv:https://doi.org/10.1080/01449290110089561, doi:10.1080/01449290110089561.
URL https://doi.org/10.1080/01449290110089561

[34] A. Sears, J. A. Jacko, J. Chu, F. Moro, The role of visual search in the design of effective soft keyboards, Behaviour & Information Technology 20 (3) (2001) 159–166. arXiv:https://doi.org/10.1080/01449290110049790, doi:10.1080/01449290110049790.
URL https://doi.org/10.1080/01449290110049790

[35] M. Kjærup, M. B. Skov, P. A. Nielsen, J. Kjeldskov, J. Gerken, H. Reiterer, Longitudinal Studies in HCI Research: A Review of CHI Publications From 1982–2019, Springer International Publishing, Cham, 2021, Ch. Theoretical Perspectives, pp. 11–39. doi:10.1007/978-3-030-67322-2_2.
URL https://doi.org/10.1007/978-3-030-67322-2_2

[36] I. S. MacKenzie, R. W. Soukoreff, Text entry for mobile computing: Models and methods,theory and practice, Human–Computer Interaction 17 (2-3) (2002) 147–198. arXiv:https://www.tandfonline.com/doi/pdf/10.1080/07370024.2002.9667313, doi:10.1080/07370024.2002.9667313.
URL https://www.tandfonline.com/doi/abs/10.1080/07370024.2002.9667313

[37] C. Yu, Y. Gu, Z. Yang, X. Yi, H. Luo, Y. Shi, Tap, dwell or gesture? exploring head-based text entry techniques for hmds, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 4479–4488. doi:10.1145/3025453.3025964.
URL https://doi.org/10.1145/3025453.3025964

[38] I. S. Mackenzie, S. X. Zhang, R. W. Soukoreff, Text entry using soft keyboards, Behaviour & Information Technology 18 (4) (1999) 235–244. arXiv:https://doi.org/10.1080/014492999118995, doi:10.1080/014492999118995.
URL https://doi.org/10.1080/014492999118995

[39] M. Silfverberg, I. S. MacKenzie, P. Korhonen, Predicting text entry speed on mobile phones, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '00, Association for Computing Machinery, New York, NY, USA, 2000, p. 9–16. doi:10.1145/332040.332044.
URL https://doi.org/10.1145/332040.332044

[40] A. Kurauchi, W. Feng, A. Joshi, C. Morimoto, M. Betke, Eyeswipe: Dwell-free text entry using gaze paths, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 1952–1956. doi:10.1145/2858036.2858335.
URL https://doi.org/10.1145/2858036.2858335

[41] L. Magnien, J. L. Bouraoui, N. Vigouroux, Mobile text input with soft keyboards: Optimization by means of visual clues, in: S. Brewster, M. Dunlop (Eds.), Mobile Human-Computer Interaction - MobileHCI 2004, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 337–341. doi:10.1007/978-3-540-28637-$0_3$3.

[42] D. Grüneis, M. Kurz, E. Sonnleitner, Let me help you: Improving the novice experience of high-performance keyboard layouts with visual clues, Applied Sciences 13 (16) (2023). doi:10.3390/app13169391.
URL https://www.mdpi.com/2076-3417/13/16/9391

[43] M. Goel, A. Jansen, T. Mandel, S. N. Patel, J. O. Wobbrock, Contexttype: Using hand posture information to improve mobile touch screen text entry, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 2795–2798. doi:10.1145/2470654.2481386.
URL https://doi.org/10.1145/2470654.2481386

[44] S. Azenkot, S. Zhai, Touch behavior with different postures on soft smartphone keyboards, in: Proceedings of the 14th International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI '12, Association for Computing Machinery, New York, NY, USA, 2012, p. 251–260. doi:10.1145/2371574.2371612.
URL https://doi.org/10.1145/2371574.2371612

[45] A. Gunawardana, T. Paek, C. Meek, Usability guided key-target resizing for soft keyboards, in: Proceedings of the 15th International Conference on Intelligent User Interfaces, IUI '10, Association for Computing Machinery, New York, NY, USA, 2010, p. 111–118. doi:10.1145/1719970.1719986.
URL https://doi.org/10.1145/1719970.1719986

[46] A. Arif, M. Pahud, K. Hinckley, B. Buxton, Experimental study of stroke shortcuts for a touchscreen keyboard with gesture-redundant keys removed, in: Proceedings of Graphics Interface 2014, GI 2014, Canadian Human-Computer Communications Society, Toronto, Ontario, Canada, 2014, pp. 43–50.
URL https://graphicsinterface.org/proceedings/gi2014/gi2014-6/

[47] O. Alsharif, T. Ouyang, F. Beaufays, S. Zhai, T. Breuel, J. Schalkwyk, Long short term memory neural network for keyboard gesture decoding, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 2076–2080. doi:10.1109/ICASSP.2015.7178336.

[48] S. Zhai, P. O. Kristensson, The word-gesture keyboard: Reimagining keyboard interaction, Commun. ACM 55 (9) (2012) 91–101.

doi:10.1145/2330667.2330689.
URL https://doi.org/10.1145/2330667.2330689

[49] S. Reyal, S. Zhai, P. O. Kristensson, Performance and user experience of touchscreen and gesture keyboards in a lab setting and in the wild, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15, Association for Computing Machinery, New York, NY, USA, 2015, p. 679–688. doi:10.1145/2702123.2702597.
URL https://doi.org/10.1145/2702123.2702597

[50] X. Bi, C. Chelba, T. Ouyang, K. Partridge, S. Zhai, Bimanual gesture keyboard, in: Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology, UIST '12, Association for Computing Machinery, New York, NY, USA, 2012, p. 137–146. doi:10.1145/2380116.2380136.
URL https://doi.org/10.1145/2380116.2380136

[51] A. Markussen, M. R. Jakobsen, K. Hornbæk, Vulture: A mid-air word-gesture keyboard, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '14, Association for Computing Machinery, New York, NY, USA, 2014, p. 1073–1082. doi:10.1145/2556288.2556964.
URL https://doi.org/10.1145/2556288.2556964

[52] X. Bi, B. A. Smith, S. Zhai, Quasi-qwerty soft keyboard optimization, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10, Association for Computing Machinery, New York, NY, USA, 2010, p. 283–286. doi:10.1145/1753326.1753367.
URL https://doi.org/10.1145/1753326.1753367

[53] M. Dunlop, J. Levine, Multidimensional pareto optimization of touchscreen keyboards for speed, familiarity and improved spell checking, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12, Association for Computing Machinery, New York, NY, USA, 2012, p. 2669–2678. doi:10.1145/2207676.2208659.
URL https://doi.org/10.1145/2207676.2208659

[54] S. Zhai, P. O. Kristensson, Interlaced qwerty: Accommodating ease of visual search and input flexibility in shape writing, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '08, Association for Computing Machinery, New York, NY, USA, 2008, p. 593–596.

doi:10.1145/1357054.1357149.
URL https://doi.org/10.1145/1357054.1357149

[55] A. Pavlovych, W. Stuerzlinger, Model for non-expert text entry speed on 12-button phone keypads, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '04, Association for Computing Machinery, New York, NY, USA, 2004, p. 351–358. doi:10.1145/985692.985737.
URL https://doi.org/10.1145/985692.985737

[56] A. U. Batmaz, A. K. Mutasim, M. Malekmakan, E. Sadr, W. Stuerzlinger, Touch the wall: Comparison of virtual and augmented reality with conventional 2d screen eye-hand coordination training systems, in: 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), IEEE, 2020, pp. 184–193. doi:10.1109/VR46266.2020.00037.

[57] A. Mutasim, A. U. Batmaz, M. Hudhud Mughrabi, W. Stuerzlinger, Performance analysis of saccades for primary and confirmatory target selection, in: Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology, VRST '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 1–12. doi:10.1145/3562939.3565619.
URL https://doi.org/10.1145/3562939.3565619

[58] A. K. Mutasim, A. U. Batmaz, W. Stuerzlinger, Pinch, click, or dwell: Comparing different selection techniques for eye-gaze-based pointing in virtual reality, in: ACM Symposium on Eye Tracking Research and Applications, ETRA '21 Short Papers, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1–7. doi:10.1145/3448018.3457998.
URL https://doi.org/10.1145/3448018.3457998

[59] I. Schuetz, T. S. Murdison, K. J. MacKenzie, M. Zannoli, An explanation of fitts' law-like performance in gaze-based selection tasks using a psychophysics approach, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 1–13. doi:10.1145/3290605.3300765.
URL https://doi.org/10.1145/3290605.3300765

[60] M. Choe, Y. Choi, J. Park, H. K. Kim, Comparison of gaze cursor input methods for virtual reality devices, International Journal of Human–Computer Interaction 35 (7) (2019) 620–629. arXiv:https://doi.org/10.1080/10447318.2018.1484054,

doi:10.1080/10447318.2018.1484054.
URL https://doi.org/10.1080/10447318.2018.1484054

[61] I. S. MacKenzie, R. W. Soukoreff, Phrase sets for evaluating text entry techniques, in: CHI '03 Extended Abstracts on Human Factors in Computing Systems, CHI EA '03, Association for Computing Machinery, New York, NY, USA, 2003, p. 754–755. doi:10.1145/765891.765971.
URL https://doi.org/10.1145/765891.765971

[62] A. S. Arif, W. Stuerzlinger, Predicting the cost of error correction in character-based text entry technologies, in: SIGCHI Conference on Human Factors in Computing Systems, CHI '10, 2010, pp. 5–14. doi:10.1145/1753326.1753329.
URL https://doi.org/10.1145/1753326.1753329

[63] A. S. Arif, W. Stuerzlinger, Analysis of text entry performance metrics, in: 2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH), IEEE, 2009, pp. 100–105. doi:10.1109/TIC-STH.2009.5444533.

[64] R. W. Soukoreff, I. S. MacKenzie, Measuring errors in text entry tasks: An application of the levenshtein string distance statistic, in: CHI '01 Extended Abstracts on Human Factors in Computing Systems, CHI EA '01, Association for Computing Machinery, New York, NY, USA, 2001, p. 319–320. doi:10.1145/634067.634256.
URL https://doi.org/10.1145/634067.634256

[65] R. W. Soukoreff, I. S. MacKenzie, 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, CHI '03, Association for Computing Machinery, New York, NY, USA, 2003, p. 113–120. doi:10.1145/642611.642632.
URL https://doi.org/10.1145/642611.642632

[66] R. W. Soukoreff, I. S. Mackenzie, Theoretical upper and lower bounds on typing speed using a stylus and a soft keyboard, Behaviour & Information Technology 14 (6) (1995) 370–379. arXiv:https://doi.org/10.1080/01449299508914656, doi:10.1080/01449299508914656.
URL https://doi.org/10.1080/01449299508914656

[67] V. Rajanna, J. P. Hansen, Gaze typing in virtual reality: Impact of keyboard design, selection method, and motion, in: Proceedings of the 2018 ACM Symposium on Eye Tracking Research and Applications, ETRA '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 1–10. doi:10.1145/3204493.3204541.
URL https://doi.org/10.1145/3204493.3204541

[68] W. Cui, R. Liu, Z. Li, Y. Wang, A. Wang, X. Zhao, S. Rashidian, F. Baig, I. Ramakrishnan, F. Wang, X. Bi, Glancewriter: Writing text by glancing over letters with gaze, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1–13. doi:10.1145/3544548.3581269.
URL https://doi.org/10.1145/3544548.3581269

[69] M. Zhao, A. M. Pierce, R. Tan, T. Zhang, T. Wang, T. R. Jonker, H. Benko, A. Gupta, Gaze speedup: Eye gaze assisted gesture typing in virtual reality, in: Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 595–606. doi:10.1145/3581641.3584072.
URL https://doi.org/10.1145/3581641.3584072

[70] V. V. Bochkarev, A. V. Shevlyakova, V. D. Solovyev, The average word length dynamics as an indicator of cultural changes in society, Social Evolution and History 14 (2) (2015) 153–175.
URL https://arxiv.org/abs/1208.6109

[71] J. F. Hair Jr, W. C. Black, B. J. Babin, R. E. Anderson, Multivariate data analysis (2014).

[72] P. Mallery, D. George, SPSS for Windows step by step: a simple guide and reference, Pearson, 2003.

[73] J. O. Wobbrock, L. Findlater, D. Gergle, J. J. Higgins, The aligned rank transform for nonparametric factorial analyses using only anova procedures, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11, ACM, New York, NY, USA, 2011, pp. 143–146. doi:10.1145/1978942.1978963.
URL http://doi.acm.org/10.1145/1978942.1978963

[74] A. M. Treisman, G. Gelade, A feature-integration theory of attention, Cognitive Psychology 12 (1) (1980) 97–136. doi:https://doi.org/10.1016/0010-0285(80)90005-5.
URL https://www.sciencedirect.com/science/article/pii/0010028580900055

[75] M. D'Zmura, Color in visual search, Vision Research 31 (6) (1991) 951–966. doi:https://doi.org/10.1016/0042-6989(91)90203-H.
URL https://www.sciencedirect.com/science/article/pii/004269899190203H

[76] A. M. Feit, D. Weir, A. Oulasvirta, How we type: Movement strategies and performance in everyday typing, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 4262–4273. doi:10.1145/2858036.2858233.
URL https://doi.org/10.1145/2858036.2858233

[77] J. Dudley, H. Benko, D. Wigdor, P. O. Kristensson, Performance envelopes of virtual keyboard text input strategies in virtual reality, in: 2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), 2019, pp. 289–300. doi:10.1109/ISMAR.2019.00027.

[78] R. S. Masters, J. P. Maxwell, Implicit motor learning, reinvestment and movement disruption: What you don't know won't hurt you, in: Skill Acquisition in Sport, Routledge, 2004, pp. 207–228.
URL https://api.semanticscholar.org/CorpusID:142428042

[79] K. Caine, Local standards for sample size at CHI, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 981–992. doi:10.1145/2858036.2858498.
URL https://doi.org/10.1145/2858036.2858498

[80] S. K. Card, T. P. Moran, A. Newell, The keystroke-level model for user performance time with interactive systems, Commun. ACM 23 (7) (1980) 396–410. doi:10.1145/358886.358895.
URL https://doi.org/10.1145/358886.358895

[81] I. S. MacKenzie, Human-computer interaction: An empirical research perspective, Newnes, 2012. doi:10.1016/B978-0-12-405865-1.00007-8.
URL https://doi.org/10.1016/B978-0-12-405865-1.00007-8

[82] A. D. Rice, J. W. Lartigue, Touch-level model (TLM): Evolving KLM-GOMS for touchscreen and mobile devices, in: Proceedings of the 2014 ACM Southeast Regional Conference, ACM SE '14, Association for Computing Machinery, New York, NY, USA, 2014, pp. 1–6. doi:10.1145/2638404.2638532.
URL https://doi.org/10.1145/2638404.2638532

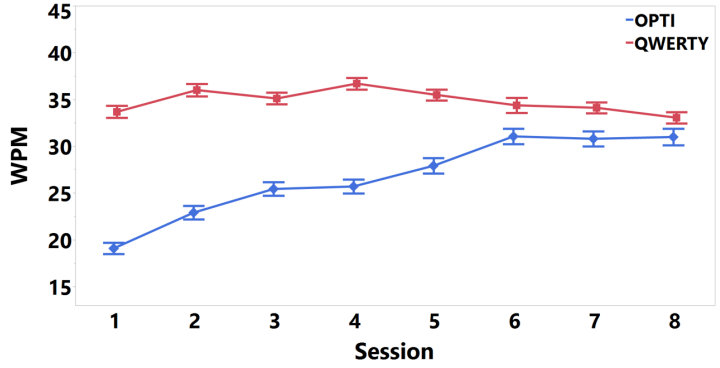
[83] J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, Y. Qin, An integrated theory of the mind, Psychological Review 111 (4) (2004) 1036–1060. doi:10.1037/0033-295X.111.4.1036.
URL https://doi.org/10.1037/0033-295X.111.4.1036

[84] A. Das, W. Stuerzlinger, A cognitive simulation model for novice text entry on cell phone keypads, in: 14th European Conference on Cognitive Ergonomics: Invent! Explore!, ECCE '07, 2007, pp. 141–147. doi:10.1145/1362550.1362579.
URL https://doi.org/10.1145/1362550.1362579

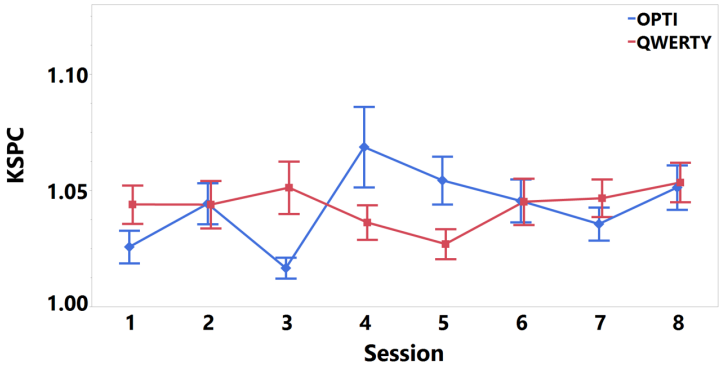
[85] A. Das, W. Stuerzlinger, Modeling learning effects in mobile texting, in: International Conference on Mobile and Ubiquitous Multimedia, MUM '08, 2008, pp. 154–161. doi:10.1145/1543137.1543169.
URL https://doi.org/10.1145/1543137.1543169

[86] A. Das, W. Stuerzlinger, Proactive interference in location learning: A new closed-form approximation, in: International Conference on Cognitive Modeling, ICCM '10, 2010, pp. 37–42.
URL https://iccm-conference.neocities.org/2010/proceedings

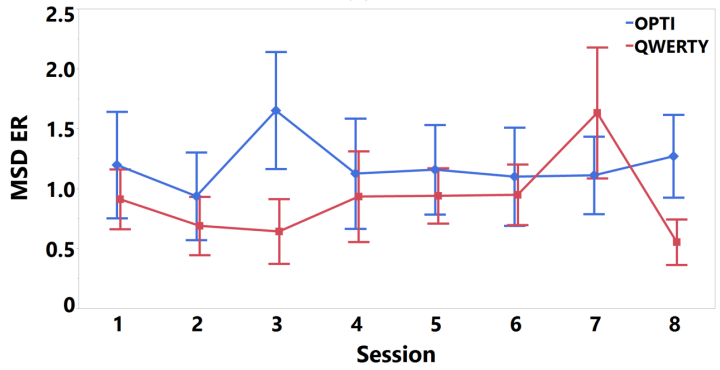
[87] A. Das, W. Stuerzlinger, Unified modeling of proactive interference and memorization effort: A new mathematical perspective within act-r theory, in: Annual Meeting of the Cognitive Science Society, CogSci '13, 2013, pp. 358–363.
URL https://escholarship.org/uc/item/1p05s7db



(a)



(b)



(c)

Figure 2: Results of Phrase Repetition Study, i.e., Study 1, for (a) WPM, (b) KSPC, and (c) MSD ER over eight sessions of typing the same phrase. The error bars show the standard error of means.

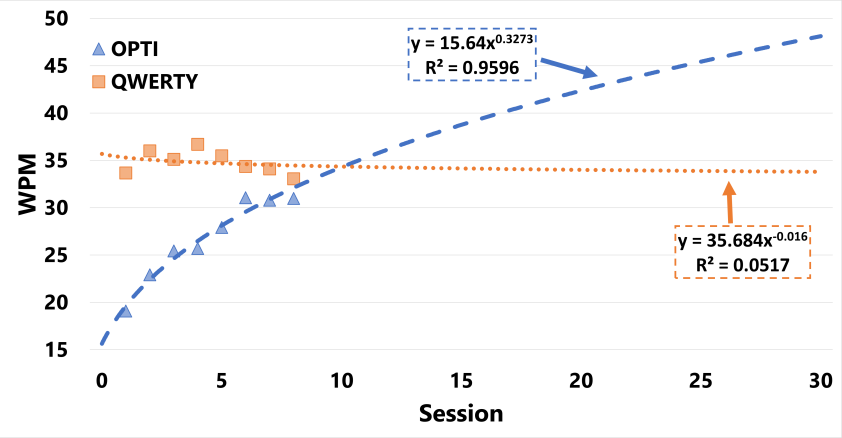


Figure 3: WPM in Phrase Repetition Study, i.e., Study 1, over eight sessions along with an extrapolation of the learning curve to the 30^{th} session.

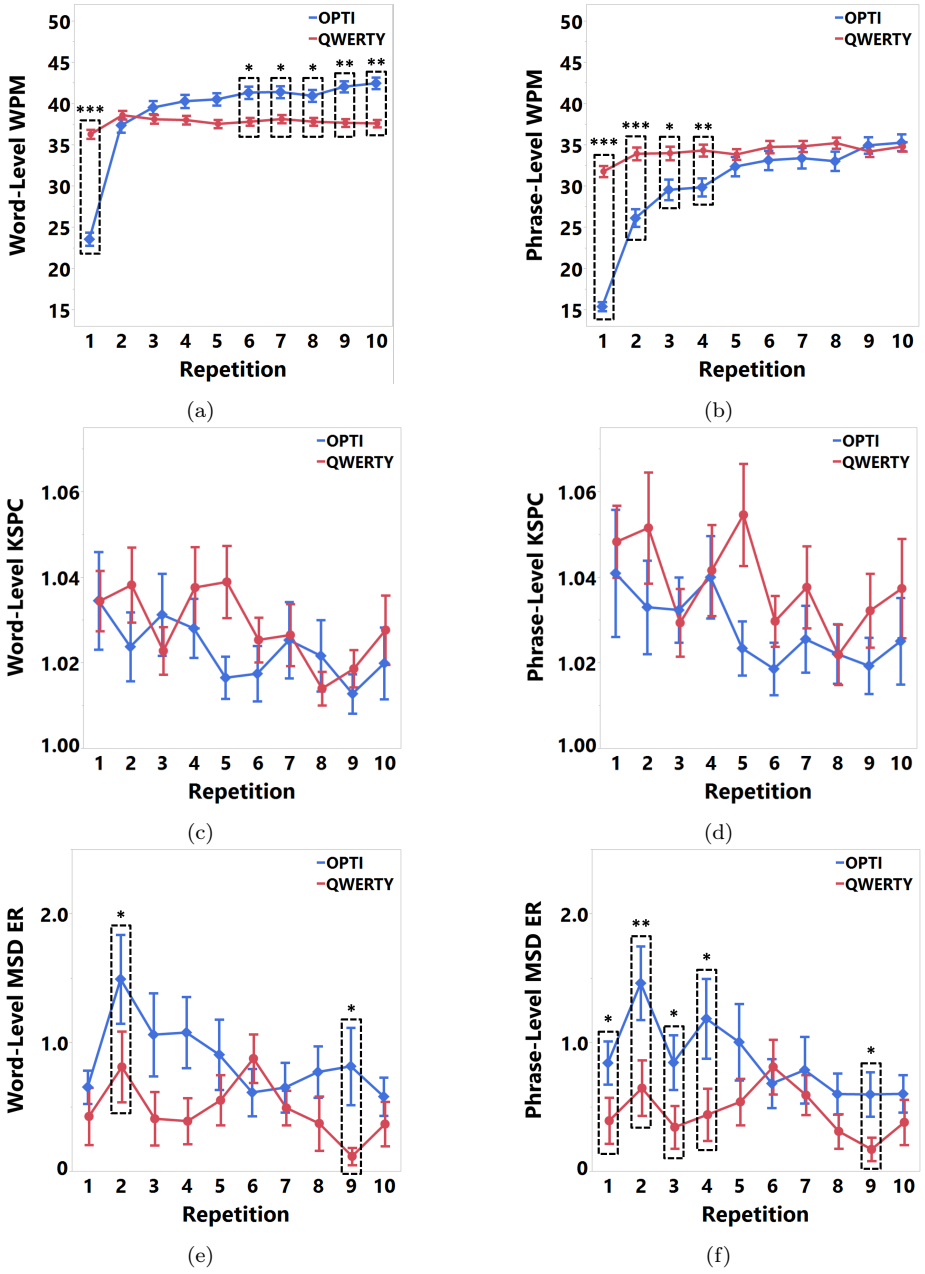


Figure 4: Results of Word Repetition Study, i.e., Study 2, for **word-level** (a) WPM, (c) KSPC, and (e) MSD ER, and **phrase-level** (b) WPM, (d) KSPC, and (f) MSD ER over ten repetitions of typing the same word. Significance levels are shown as *** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$. The error bars show the standard error of means.

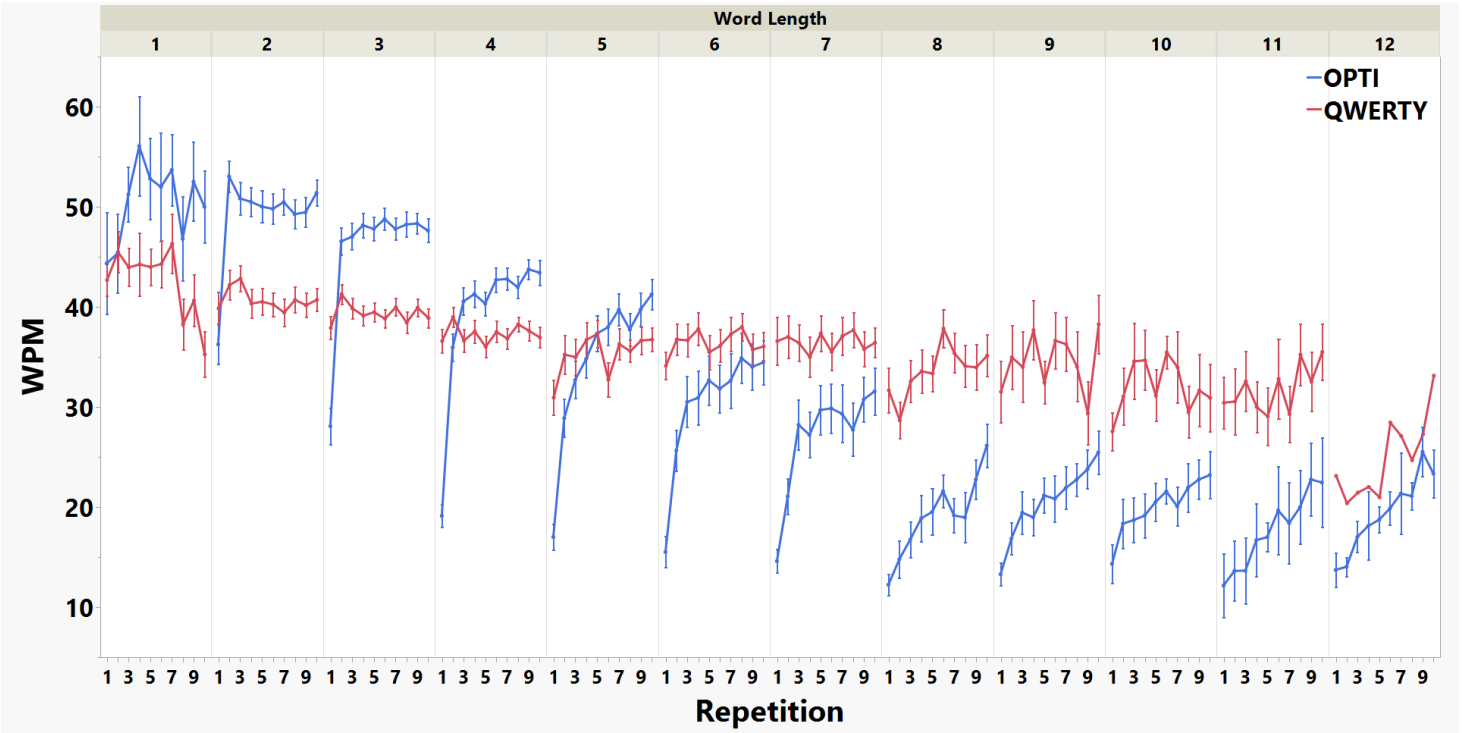
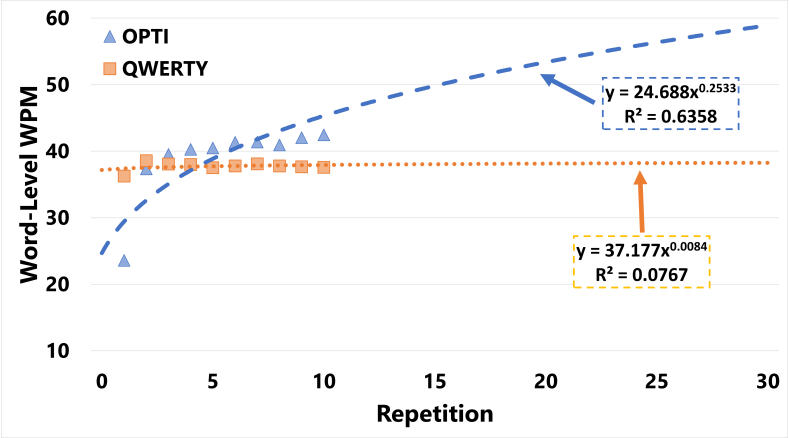
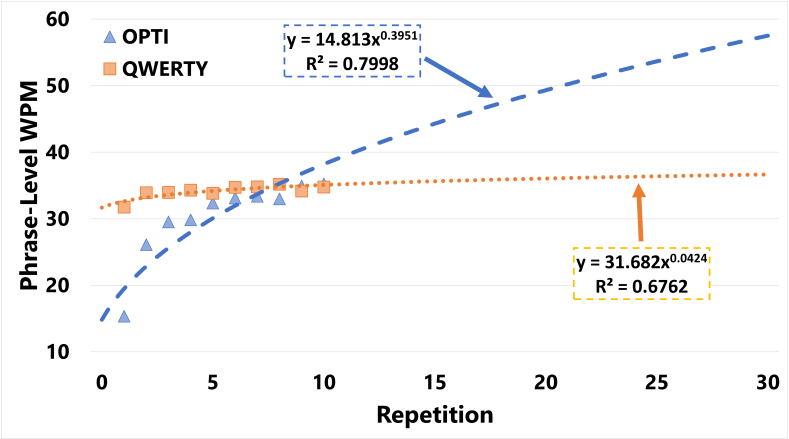


Figure 5: WPM in Word Repetition Study, i.e., Study 2, for each Word Length over ten repetitions. The error bars show the standard error of means.



(a)



(b)

Figure 6: (a) **Word-level** and (b) **phrase-level** WPM in Word Repetition Study, i.e., Study 2, over ten repetitions along with an extrapolation of the learning curve to the 30^{th} repetition.

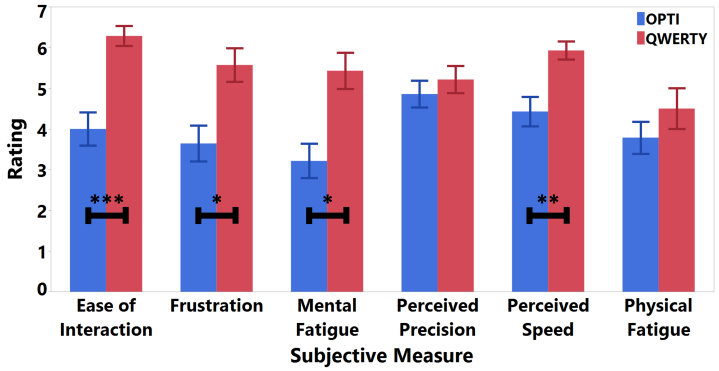


Figure 7: Results of Word Repetition Study, i.e., Study 2, for Subjective Measures of the OPTI and QWERTY keyboards. Significance levels are shown as *** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$. The error bars show the standard error of means.

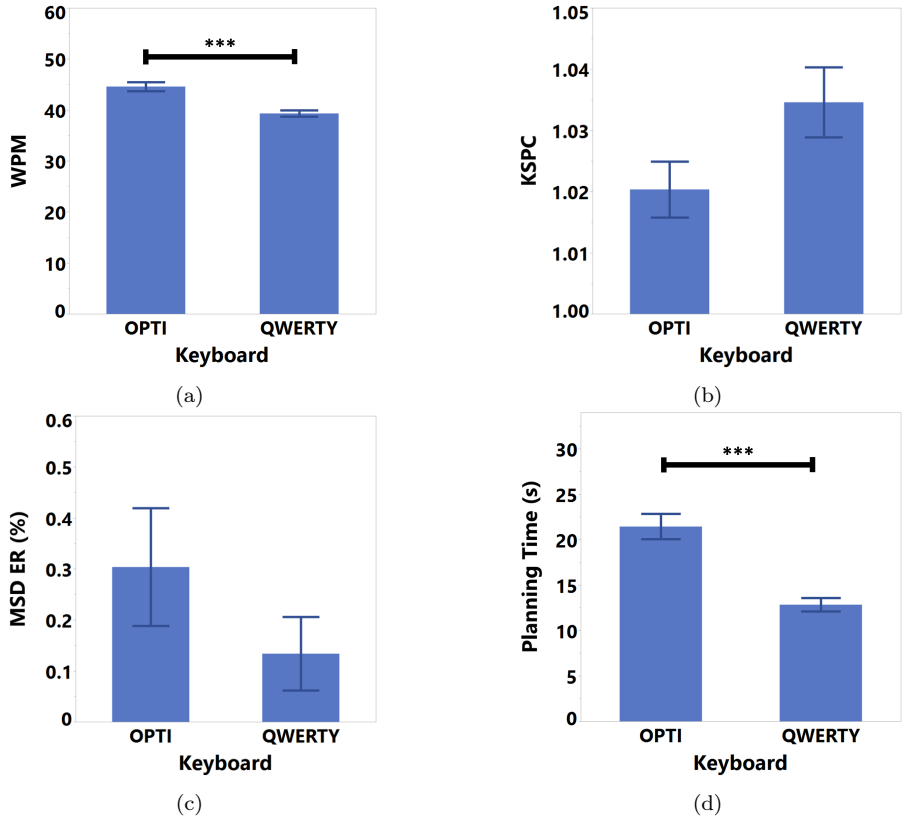
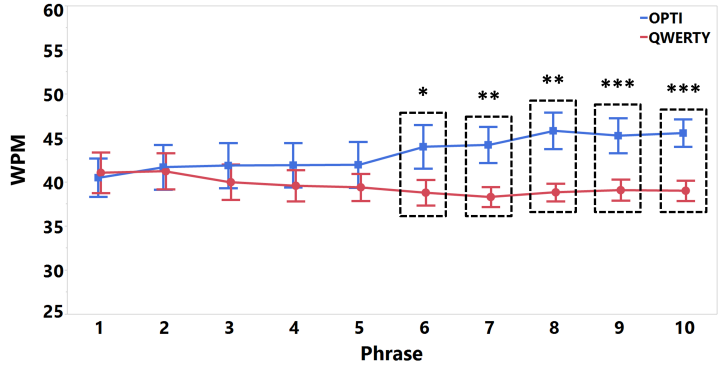
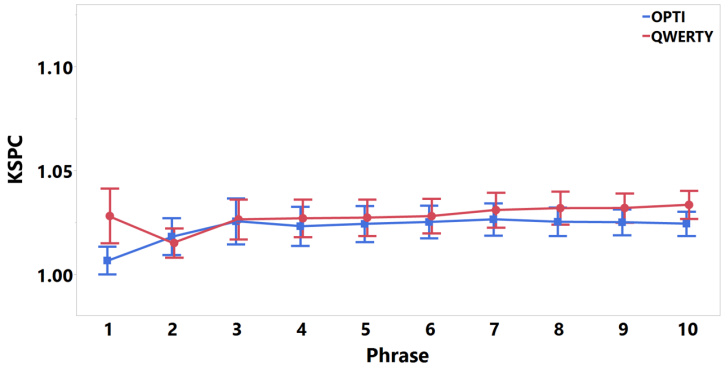


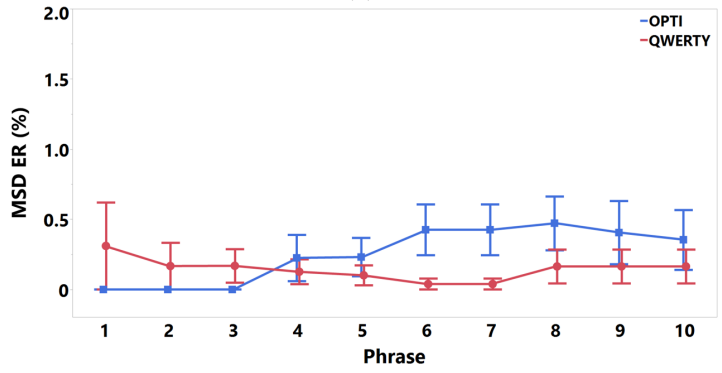
Figure 8: Results of GEM Study, i.e., Study 3, for overall average (a) WPM, (b) KSPC, (c) MSD ER, and (d) Planning Time of all the phrases for OPTI and QWERTY. Significance levels are shown as *** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$. The error bars show the standard error of means.



(a)

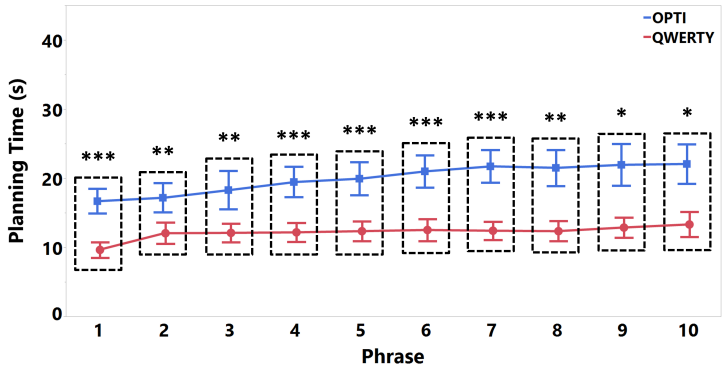


(b)

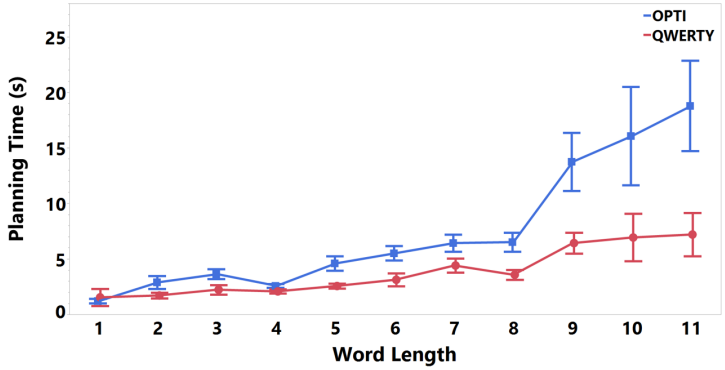


(c)

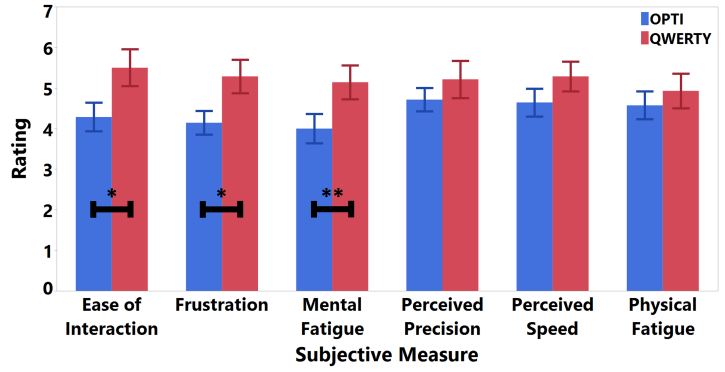
Figure 9: Results of GEM Study, i.e., Study 3, for 5-phrases Moving Average of (a) WPM, (b) KSPC, and (c) MSD ER. Significance levels are shown as *** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$. The error bars show the standard error of means.



(a)



(b)



(c)

Figure 10: Results of GEM Study, i.e., Study 3, for (a) 5-phrases Moving Average of Planning Time, (b) Planning Time for each Word Length, and (c) Subjective Measures for OPTI and QWERTY keyboards. Significance levels are shown as *** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$. The error bars show the standard error of means.