Improving Novice Performance on the Twiddler One–Handed Chording Keyboard

Kent Lyons[1], Brian Gane[2], Thad Starner[1], Richard Catrambone[2]

[1] College of Computing
[2] School of Psychology
Georgia Institute of Technology
Atlanta, GA 30332 USA
{kent[1], thad[1]}@cc.gatech.edu
{gtg678s[2], rc7[2]}@prism.gatech.edu

Abstract. The Twiddler one–handed chording keyboard offers rapid mobile text entry on a keypad similar to that of a standard mobile telephone. Previously we found that novice users could be trained to type 47 words per minute (wpm) while one participant reached 67 wpm. Here, we present an evaluation designed to examine which combination of two typing aids helps novice Twiddler typists the most. Specifically, we examine the effects of a phrase set designed for the Twiddler and the manipulation of an on–screen keymap representation. Sixty participants were divided across 6 conditions and typed for two 20 minute sessions. We found that, where there is an effect, using our ordered phrase set aids novice Twiddler typists' typing rate, error rate and mental workload. Highlighting our on–screen representation helps typing speed, accuracy, and reduces workload.

1 Introduction

Mobile computing is becoming one of the most widely adopted computing technologies. As of 2004, there are 1.3 billion mobile phone subscribers and there could be as many as 2 billion by 2007 [1]. Wireless text messaging is widespread with predictions of a rate of over 1 trillion messages per year being reached shortly [7, 13]. Unfortunately, slow text entry on mobile devices may limit the utility of upcoming services such as wireless email.

We have found that the Twiddler chording keyboard (Figure 1) offers very rapid mobile text entry rates [9, 8]. We believe that this keyboard is a viable alternative for mobile devices because it employs the same 3 x 4 button layout as a mobile phone. While this keyboard outperforms many other methods, we found that the initial typing rate is slower than some other methods. In this paper, we explore whether typing aids can increase novice typing rates and reduce the barriers to Twiddler acceptance.

2 Typing on Mobile Phone Keypads

There are two ways to accommodate the small form-factor keyboards that are resulting from the decrease in size of mobile technology: make the keys very small, like on mini-QWERTY keyboards, or remove the one–to–one mapping between keys and characters. Most phones map more than one character onto a key because they inherited the

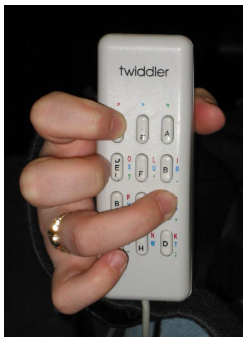

Fig. 1. Chord for the letter ‘j’ on the Twiddler

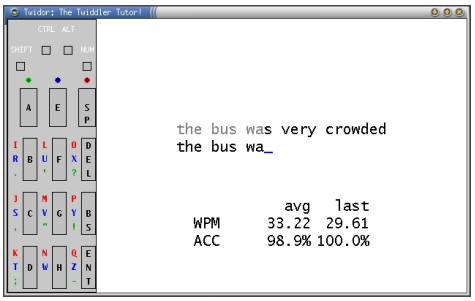

Fig. 2. Experimental software showing the chording layout, phrase and statistics.

12 button keypad of traditional phones. When multiple characters are assigned to one key, a method is needed to disambiguate between the possible options. Wigdor and Balakrishnan [16] present a taxonomy with three dimensions for ways to disambiguate: the number of keys used (one or more), the number of presses performed on the key(s) and the possible temporal ordering of key presses (consecutive or concurrent). These methods can be further combined with linguistic models to disambiguate the key presses.

Multi–tap is a very common text entry technique for mobile phones. The alphabet is mapped onto 8 of the 12 buttons on the mobile phone keypad resulting in 3 or 4 letters per key. To generate a character and disambiguate between the characters on the same key, the user presses a single key multiple times to cycle through the letters until the desired one appears on the screen. While a very common typing method, it is also relatively slow. Previous research has found multi–tap typing rates for novice users ranging from 7.2–8.7 wpm with 15–30 minutes of practice [10, 15, 16, 9]. These studies show that as users gain experience their typing rates can increase to 11.0–19.8 wpm.

T9 is another common mobile phone input method. Like multi–tap, multiple letters are assigned to each button on the keypad. However instead of having the user disambiguate every character with multiple button presses, T9 uses language disambiguation. Using a dictionary, T9 presents the most probable string the user is attempting to enter given the input so far. If the presented text is incorrect, the user can press a special key to cycle through possible alternatives. One study found that novice users type 9.1 wpm while experts can achieve 20.4 wpm [5]. Unfortunately, T9 rates drop drastically once the user needs to enter words that are not in the dictionary.

Recently there has been a resurgence in mobile phone keyboard entry research which has produced a number of new methods to enter text on the mobile phone keypads including LetterWise [10], TiltText [15], and ChordTap [16]. These methods offer novice performance similar to multi–tap (7.3 wpm, 7.4 wpm and 8.5 wpm respectively). In addition, each of these methods offers faster expert typing rates than multi–tap given the same amount of practice. LetterWise users achieved a rate of 21 wpm after approximately 550 minutes of practice. TiltText users reached 13.6 wpm and ChordTap 16.1 wpm respectively with about 160 minutes of typing practice.

3 Twiddler Chording

Unlike a mobile phone, the Twiddler is held with the keypad facing away from the user (Figure 1), and each row of keys is operated by one of the user's four fingers. Instead of pressing keys in sequence to produce a character, multiple keys can be pressed simultaneously to generate a chord. The default keymap for the Twiddler consists of single button and two button chords which are assigned in an alphabetical order and is divided into three parts (Figure 3). Characters 'a'–'h' only require one button press ("single"). The letters 'i'–'q' and 'r'–'z' are typed with chords of two buttons. For these letters, two of the buttons on the top row act as shift keys. The shift button for 'i'-'q' is called the red shift, and the shift for 'r'–'z' is the blue shift. This nomenclature is derived from the keymap printed on the face of the Twiddler.

Previously, we evaluated the relative learning rates of typing with multi–tap versus typing with chording on the Twiddler [9]. We conducted a longitudinal study with ten participants which had no experience with typing chords on the Twiddler and varying levels of practice typing with multi–tap. The experiment was a within–subjects design; participants typed in two conditions (multi–tap and chording) during 20 sessions of typing. Each session consisted of two parts delineated by typing condition and a five minute break in the middle. Our experimental software (Figure 2) prompts the participant with the phrase to be typed (selected randomly from the MacKenzie and Soukoreff phrase set [11]) and records the response and timings for all of the buttons pressed.

We found the mean entry rates for our ten participants for session one were 8.2 wpm for multi–tap and 4.3 wpm for chording. As sessions continued, the means improved and reached 19.8 wpm for multi–tap and 26.2 wpm for chording after 20 sessions or 400 minutes of practice. While both conditions showed improvement, the typing rates for the chording condition rapidly surpassed those of multi–tap and chording showed strong signs of continued learning. A regression analysis predicted that the faster typists would reach 60 wpm after 10,000 phrases (approximately 80 sessions or 27 hours) while the slower typists could achieve 45 wpm.

Our second study on the Twiddler was designed to confirm the predictions of our regression curves for expert rates [8]. Five of our ten participants agreed to continue in our followup experiment. The five that declined to continue participating did so because of the large additional time commitment. The procedure was modified to focus on chording; we replaced the multi–tap condition from our original experiment with a second chording session. By the end of the study, each of our participants completed an average of 75 sessions which corresponds to approximately 25 total hours of practice. On average, our participants reached a typing rate of 47 wpm. Surprisingly, one subject achieved a rate of 67.1 wpm, equivalent to the typing rate of the third author, an expert who has been a Twiddler user for ten years.

While the Twiddler shows great potential for permitting rapid text entry in a mobile environment, our studies did show that the initial typing rate was about half that of multi–tap. The crossover point, where the chording typing rate overtook multi–tap, occurred after the fifth session or after 100 minutes of practice. In this paper, we explore ways to improve novice typing rates and acceptance of the Twiddler.

4 Aiding Novice Twiddler Typing

The orientation of the hand while typing on the Twiddler is more like a musical instrument such as a guitar than a computer keyboard. While offering good expert rates, this orientation poses a problem for novices; it makes "hunt–and–peck" typing difficult. To look at which key to press, a user needs to rotate the Twiddler out of typing position to bring the keypad into view. The second potential barrier for novice users is chording, pressing multiple buttons simultaneously to generate a character. While chording is employed on desktop keyboards (shift, control and alt are often used as one button of a chord) it is more rare on mobile phone keypads. Furthermore for the Twiddler, the majority of the characters in the alphabet require the use of chording. To address these potential problems, we are exploring two aids that might aid novice users: a structured phrase set and software highlighting for the keys to be pressed.

4.1 Phrase Set

Our first aid employs a phrase set tailored to the Twiddler keymap. One common practice with tutors for desktop keyboards is to subdivide the alphabet based on the physical layout of the keyboard. For instance, the software starts by teaching the user the "home row" and gradually adds more letters to be learned based on the position of the keys on the keyboard. We extend this analogy to the Twiddler keymap and have different phrases that exercise different categories of chords. Our new phrase set is initially restricted so that the user only type letters requiring a single button press ('a'–'h'). Next the phrase set is changed so the participant types just the chords that involve the red shift ('i'–'q'). Then, the phrase set uses the combination of single and red ('a'-'q'), followed by just blue ('r'-'z'), single and blue, and finally all of the letters.

In addition, empirical evidence from psychology studies indicates that simplifying a complex task into smaller tasks can reduce the workload associated with learning the complex task and can reduce error rates [2, 6]. Our new phrase set can be ordered so that the task of learning all 26 letters of the alphabet is simplified into several subtasks. Each task focuses on learning subsets of the alphabet where each subset is associated with a critical gross physical movement. By segmenting the phrase set based on the different types of chords, we can help the user focus on the different types of physical movements needed to type. The phrases that use only a single button let the user explore the keyboard. The red and blue phrases give practice for the motions needed to type the different chords involving the two shift keys. Finally, the phrases which use combinations start to transition the user to more realistic text and the associated movements required.

4.2 Highlighting

Our second aid supplements an on–screen keyboard representation which provides the user a reference of the mapping between buttons and characters (Figure 3). The representation is shown to the user on the left–hand portion of the display (Figure 2) and is the same as the representation which is printed on the faceplate of the Twiddler. All of the characters for single button chords are printed on the button. The characters for the rest of the chords are printed in their respective colors next to the appropriate button.

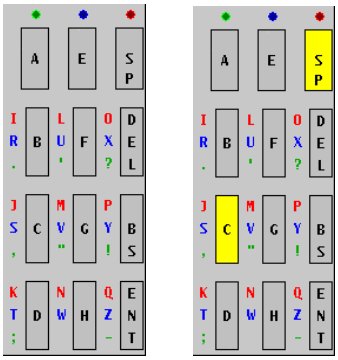

Fig. 3. Graphical representation of Twiddler chording keymap. Shown without highlighting (left) and with (right).

We provided this on–screen representation in our previous studies so that our participants could use it as a reference while learning to type. It is designed to help reduce the need to turn the Twiddler in order to look at the keypad. Instead, participants can scan the on–screen representation to find the letter they need to type. Once the correct letter is found, the participants can determine which buttons to press. While informative, it is visually busy and requires some experience to understand and use. To facilitate the use of the Twiddler representation, our software can highlight the next set of buttons the user is to press (Figure 3, right). The highlighting is designed to reduce the amount of time the user spends visually scanning the representation. When highlighting is turned on, the buttons to be pressed for the next character change color.

Next, we present our study designed to explore these two aids. Our goal is to determine if either aid can improve novice Twiddler typing and see which combination can lead to the best novice typing rates and least workload.

5 Experiment

Our experiment retains the same core design from our previous Twiddler studies [9, 8] which were based on other text entry research [10, 12]. For these studies, experimental software presents a sequence of phrases one at a time and the participants are asked to type the displayed text. Phrases are grouped into twenty minute typing sessions and the experimental variables can be manipulated per session.

5.1 Design

We are using two twenty minute sessions in this experiment: practice and evaluation. For the practice session, we manipulate our experimental variables across participants while the evaluation session is the same for all participants. Our first variable corresponds to the phrase set aid. Our Twiddler phrase set has 14 phrases that require only single button presses, 14 phrases that only require the red shift and 14 of blue. We have

Characters	Example Phrase
Red	i look ill in pink
Single	dad added a facade
Single + Red	a feminine chief in old age
Blue	suzy trusts wussy russ
Single + Blue	the greatest war there ever was

Table 1. Example phrases exercising different portions of the Twiddler keymap.

26 phrases that use single and red characters and 25 that use single plus blue. Table 1 shows some example phrases from each of our categories. In total, our 93 phrases have an average length of approximately 25 characters and the correlation with the frequency of characters in English is 89% [11]. For the practice session we have two phrase set conditions. The first condition, "ordered," presents the phrases in a structured order. Initially, our software randomly selects phrases that require single button presses. Next it uses only "red" phrases, then single plus red, blue, and single plus blue. Our second condition, "unordered," randomly displays any of the phrases for the whole period. This condition allows us to control the content of phrase set but does not offer the aid of learning in sequence. For our evaluation session, we use the phrase set developed by MacKenzie and Soukoreff [11]. These phrases average approximately 28 characters each and are selected randomly from the set of 500 total phrases. The phrases contain only letters and spaces, and we altered the phrases to use only lower case and American English spellings. These are phrases specifically designed as representative samples of the English language and have a correlation with English of 95%.

Our second variable is highlighting, and we are testing three highlighting modes during the practice session: no highlight, always on highlight, and delayed highlight. For no highlighting, the on–screen representation is shown but does not change. When highlighting is turned on, the buttons for the next character to be typed are highlighted in yellow (Figure 3). Our software also has a delayed highlighting option, in which initially no buttons are highlighted. In this case, no prompt is shown initially. If there is no activity, the keys to press are highlighted following a delay. After pilot testing a 1.5s delay was chosen. This value was large enough to allow the pilot subjects to type many of the characters they had already learned without the highlight appearing. This value also corresponds to typing at 8 wpm which, as discussed previously, is the rate at which many novices type with other mobile phone methods. For the practice session, each participant is assigned to one of the three highlighting categories. For the evaluation session highlighting is turned off for all participants. As a result, our experiment is a 3 x 2 design. We have three highlighting and two phrase set possibilities resulting in a total of six between–subject conditions.

5.2 Participants

We recruited 60 students from the Institute. The majority participated in return for credit in their respective courses and a few students volunteered. As in our previous experiments, all of our participants have no experience with the Twiddler. Each participant

is assigned randomly to one of the six conditions resulting in ten participants per condition. Our participants ranged in age from 18 to 37 years old and had a mean age of 20.9 years ($SD = 3.7$). Thirty–one participants were female and four left handed. Twelve participants were non–native English speakers. The non–native speakers had been speaking English on average 8.9 years ($SD = 6.4$). Fifty–two of our participants were mobile phone owners. The owners made an average of 6.6 calls per day ($SD = 5.4$) and sent an average of 2.3 text messages each day ($SD = 4.5$).

5.3 Procedure

The experiment takes approximately 90 minutes. It begins with the researcher presenting an overview of the experiment, and consent and demographic forms are filled out. Next, each participant types using a standard desktop QWERTY keyboard for three minutes. We collected this data to obtain a baseline typing rate for each participant. Following the desktop keyboard test, the participants are given written instructions explaining how to hold and type with the Twiddler and how the typing software works. As appropriate, the instructions explain the breakdown of the phrase set and how the software highlighting works. For each segment of the study, we instruct the participants to type "as quickly and accurately as possible."

The first session of Twiddler typing starts next. The practice session starts with a warm–up round which consists of typing the two phrases, "abcd efgh ijkl" "mnop qrst uvwx yz" twice. The warm–up data is not used in measuring performance. After the warm–up, the participant begins the practice session. At that point the twenty minute timer starts and data recording begins. The practice session is divided into six blocks. If the participant is using the ordered phrase set, each block switches from one set of chords to the next. Four minutes is spent on single phrases, and four on red. This is followed by two minutes of practice using the single plus red phrases. Next is four minutes of blue and two minutes of single plus blue. Finally, there is four minutes of typing where phrases are selected randomly from the entire phrase set resulting in twenty minutes total. The unordered condition uses blocks of the same duration; however for each block, phrases are selected randomly from the entire phrase set. Once the twenty minutes of the practice session is complete, the participants take a five minute typing break. During the break they fill out a NASA Task Load Index (TLX) questionnaire [3]. The evaluation session starts once the questionnaire is completed and the break is over.

At the beginning of the evaluation session the participants are instructed that the highlighting will be turned off for the upcoming session (for those who had highlighting in the practice session). At this point, the software switches to using the MacKenzie phrase set for all participants. After typing the alphabet twice, participants resume typing. The evaluation session is divided into four blocks of five minutes. At the end of the twenty minute session, participants fill out a second NASA-TLX questionnaire based on the evaluation session only.

5.4 Software and Equipment

The testing software is self–administered under researcher supervision. It presents the participants with the key layout for chording (Figure 3) and statistics of performance

so participants can monitor their progress. A phrase is displayed on the screen, and the subject's typed text appears immediately below the presented text (Figure 2). The software has a built in scripting engine used to configure and control the experimental conditions. We have six scripts (one for each of our conditions) that are used by the software to run the participants through our experimental procedure.

The experiment is conducted as a stationary test where participants sit at a computer running the test software in our laboratory. The computer stations run our software (written in Java) and are Pentium III based PCs running Linux and the X windowing system. The Twiddler is attached to the computer via a serial cable and continually sends the state of all of its buttons to the computer at 2400 baud, resulting in a key sample rate of approximately 45Hz. The software parses the serial stream as text input. The software collects data at the level of button presses. Every key press and release is recorded to a log file. When a button is pressed or released, the system logs the time–stamp (obtained with Java's *System.currentTimeMillis()* system call), the character generated (if any), and the state of all of the Twiddler's buttons. The current text entry method is logged as well as the phrases presented to the participant. With this data, we can determine when each button was pressed and released, the duration each button was held, the time between releasing one button and pressing the next, and the resulting transcribed text.

6 Results

Across our 60 participants we collected approximately 3500 phrases of chording data which resulted in 84,000 transcribed characters. Using this data, we examine the effects of our experimental manipulations on participants' typing speed, error rate, and workload. We performed a 3 (highlighting) x 2 (phrase set) x 2 (session) ANOVA on each measure. Highlighting and phrase set are between-subject variables, while session is a within-subject variable. The inclusion of session allows us to determine the presence and magnitude of typing speed changes between the practice and evaluation sessions. Where appropriate, we also examine the individual 2–way interactions and simple effects of each manipulation. All results are interpreted using $\alpha = 0.05$.

6.1 Text Entry Rates

First, we examine the effect our conditions have on typing speed which is measured in words per minute (wpm). For each participant, we calculated the cumulative wpm value across an entire session by taking the sum of the total number of words and dividing by the total time spent typing in the session. Table 2 displays each group's mean wpm and standard deviation for both the practice and evaluation session.

There is no 3–way interaction between our variables, $F_{(2,54)} = 1.12, p = 0.34$, $MSE = 1.00$. There is a significant interaction between the highlight manipulation and session, $F_{(2,54)} = 8.43, p = 0.001, MSE = 7.58$. A simple effects analysis demonstrates that the highlighting off group typed slower in the practice session than in the evaluation session, $F_{(1,54)} = 9.32, p < 0.01$. In contrast, the highlighting on group typed faster in the practice session than in the evaluation session, $F_{(1,54)} = 7.02, p =$

Practice Session	Highlight off	delay	on	*Mean*
ordered	6.61 (2.58)	6.73 (1.20)	6.21 (1.16)	6.52 (1.72)
unordered	5.17 (1.74)	4.88 (1.15)	6.34 (1.69)	5.46 (1.63)
Mean	5.89 (2.26)	5.81 (1.49)	6.28 (1.41)	5.99 (1.75)

Evaluation Session	Highlight off	delay	on	*Mean*
ordered	6.92 (2.15)	6.61 (1.50)	5.42 (1.80)	6.32 (1.89)
unordered	6.69 (1.87)	5.69 (1.48)	5.55 (2.20)	5.98 (1.88)
Mean	6.80 (1.96)	6.15 (1.53)	5.48 (1.96)	6.15 (1.88)

Table 2. Mean typing rates in words per minute (with standard deviations) for the practice and evaluation sessions for all 6 groups.

0.01. The delay group exhibited no reliable difference in typing rate between the practice and evaluation sessions, $F_{(1,54)} = 1.33, p = 0.25$. A simple effects analysis of highlighting for each session revealed no significant differences, suggesting that there is no overall difference between the highlighting groups for the practice session, nor for the evaluation session.

There is a significant interaction between phrase set and session, $F_{(1,54)} = 4.26, p = 0.04, MSE = 3.83$. Simple effects analysis of phrase set in the practice session reveals that the ordered phrase set group typed faster than the unordered phrase set group, $F_{(1,54)} = 6.01, p = 0.02, MSE = 2.77$. In contrast, during the evaluation session there is no significant difference between phrase set groups, $F_{(1,54)} = 0.501, p = 0.48, MSE = 3.44$. Next, we examine if the typing rate changed between the practice and evaluation sessions for either phrase set group. A simple effects analysis reveals that the unordered group's rate increased from practice to evaluation, $F_{(1,54)} = 4.39, p = 0.04$. In contrast, the ordered group's rate did not differ between practice and evaluation, $F_{(1,54)} = 0.68, p = 0.41$. These results suggest that during the practice session the ordered phrase set allows faster typing than the unordered phrase set. In the evaluation session, when all participants typed using the same phrase set, there is no statistical difference in the typing rate.

6.2 Error Rates

Next, we examine the number of errors made. Table 3 shows the percent error means and standard deviations for each group. We are using Soukoreff's and Mackenzie's total error rate metric [14]. This metric accounts for both corrected and uncorrected errors made by the participants and provides a single total error rate.

There is no significant 3–way interaction, $F_{(2,54)} = 0.737, p = 0.48, MSE = 0.002$, indicating that we can analyze the data as three 2–way ANOVAs. There is no significant interaction between phrase set and highlighting, nor between phrase set and session for the participants' error rates.

As with typing rate, there is a significant interaction between highlighting and session, $F_{(2,54)} = 4.59, p = 0.01, MSE = 0.002$. Using a simple effects analysis we can determine how the highlighting manipulation changes as a function of session. Highlighting has a significant effect on error rates in the practice session, $F_{(2,54)} = 3.50, p =$

Practice Session	Highlight off	delay	on	*Mean*
ordered	19.9 (11.7)	14.5 (4.8)	12.1 (5.5)	15.5 (8.4)
unordered	13.6 (6.3)	15.7 (6.7)	10.0 (4.4)	13.1 (6.1)
Mean	16.8 (9.7)	15.1 (5.7)	11.1 (4.9)	14.3 (7.4)

Evaluation Session	Highlight off	delay	on	*Mean*
ordered	15.2 (10.0)	13.0 (8.4)	15.5 (7.5)	14.6 (8.5)
unordered	13.0 (6.9)	13.1 (3.4)	13.6 (7.9)	13.2 (6.1)
Mean	14.1 (8.4)	13.0 (6.3)	14.5 (7.5)	13.9 (7.4)

Table 3. Mean percent error (with standard deviations) for the practice and evaluation sessions per group.

$0.04, MSE = 0.005$, but not in the evaluation session, $F_{(2,54)} = 0.21, p = .82$. A post–hoc contrast reveals that in the practice session the highlighting on group made fewer errors than the other two highlighting groups, $t_{(57)} = 2.50, p = 0.02$.

Next, we examine how the highlighting manipulations impact error rates as participants move from the practice to evaluation sessions. For participants with highlighting on, error rates increase between the practice and evaluation sessions, $F_{(1,54)} = 4.85, p = 0.03$. There is no significant error rate differences between the practice and evaluation sessions for either the highlighting off group, $F_{(1,54)} = 2.88, p = 0.10$, or the delay highlighting group, $F_{(1,54)} = 1.68, p = 0.20$. This result suggests that error rates, which are significantly lower for the group with highlighting on during the practice session, increased to the level of the other highlighting groups during the evaluation session.

6.3 Workload

The NASA Task Load Index (TLX) questionnaire measures subjective workload ratings. Previous studies have indicated that it is a reliable and valid measure of the workload imposed by a task [3, 4]. Subjective workload ratings can be more sensitive to working memory demands than measures of performance [17]; this is important given the need for the participants to remember the Twiddler key mapping. Additionally, subjective ratings can be informative when a task is difficult, yet within the individual's capability. For instance, as a task becomes more difficult, the individual can increase his or her effort in order to maintain the same level of performance. In this case subjective ratings of workload could capture this increased effort, whereas performance measures could not [17].

The NASA-TLX consists of six scales: mental demand, physical demand, temporal demand, performance, effort, and frustration; each scale has 21 gradations. For each scale, individuals rate the demands imposed by the task. In addition, they rank each scale's contribution to the total workload by completing 15 pairwise comparisons between each combination of scales. This procedure allows an investigation of the task demands load on each scale, as well as a measure of the global workload.

Interpretation of the mental, physical, and temporal demand scales are straightforward; each scale captures the demand imposed by its title. The performance scale captures how successful participants felt they were at accomplishing the given task. The

effort scale captures how hard individuals had to work in order to achieve their level of performance; both mental and physical effort can contribute to this scale. The frustration scale captures how much the task annoys or discourages individuals.

The overall workload rating is calculated by summing the product of each scale's rating and weight. This calculation results in a score between 0 and 100. It reflects an individual's perception of the amount of workload devoted to each of the scales, along with each scale's contribution to overall workload [4]. Here, we analyze the overall workload ratings in addition to the six individual scale ratings. As with typing and error rates, for each analysis a 3 (highlighting) x 2 (phrase set) x 2 (session) ANOVA is used.

Overall Workload An analysis on the overall workload does not reveal any interesting effects. There is no significant main effect for highlighting, phrase set, or session. In addition, there is no significant interaction between highlighting and phrase set, highlighting and session, and phrase set and session. Finally, there is no 3–way interaction between highlighting, phrase set, and session. Although the overall workload score revealed no effects, an analysis of individual workload scales can still reveal relevant information about how the typing task contributes to different sources of workload [3]. For each scale, the rating (0–20) is analyzed, without regard to the participant's weighting of that scale. On each scale a higher rating reflects more workload or difficulty.

Physical Demand There is no significant 3–way interaction between highlighting, phrase set, and session. Moreover there is no significant interaction between highlighting and phrase set nor highlighting and session. Finally, there is no significant main effect for highlighting. However, there is a significant interaction between phrase set and session $F_{(1,54)} = 13.72, p < 0.01, MSE = 8.18$. The ordered group rated physical demand lower in the practice session $(M = 8.42, SD = 5.13)$ than the evaluation session $(M = 11.27, SD = 5.16)$, $F_{(1,54)} = 13.88, p < .01$. The unordered group did not rate physical demand differently between the practice and evaluation session. Simple effects were further examined by analyzing the effects of phrase set in the practice session and the evaluation session. In the practice session the ordered group rated physical demand significantly lower $(M = 8.42, SD = 5.13)$ than did the unordered group $(M = 12.63, SD = 5.22)$, $F_{(1,54)} = 7.56, p = 0.01, MSE = 29.41$. However, in the evaluation session no significant difference in ratings was found between the two phrase set groups. This suggests that the increase in physical demand between sessions for the ordered group is a result of demand being lowered in the practice session; in the evaluation session the physical demand was not different for either group.

Effort For the effort scale, there is no significant 3–way interaction between highlighting, phrase set, and session. Also, there is no significant interaction between highlighting and phrase set nor between phrase set and session. Furthermore, there is no significant main effect of phrase set indicating that the phrase set manipulation did not change participants' rating of the effort required to type on the Twiddler. The highlighting manipulation does interact with session, $F_{(2,54)} = 8.48, p = 0.001, MSE = 3.86$. A simple effects analysis of session at each level of highlighting reveals that the highlighting off group does not report significantly different amounts of effort between the

practice and evaluation sessions. However, the highlighting on group rated the effort required to type in the practice session lower ($M = 13.38, SD = 4.11$) than the effort required in the evaluation session ($M = 14.93, SD = 3.61$), $F_{(1,54)} = 5.64, p = 0.02$. In contrast, the delayed highlighting group reported higher effort in the practice session ($M = 13.80, SD = 3.30$) compared to the evaluation session ($M = 12.70, SD = 3.93$), $F_{(1,54)} = 11.16, p < 0.01$. Further simple effects analyses reveal that the three highlighting groups are not significantly different in either the practice or evaluation session.

Mental and Temporal Demand, Performance, and Frustration The software manipulations do not have any significant effects on mental demand ratings or performance ratings. There is only one significant difference for ratings on the temporal demand scale. There is a significant main effect for session, indicating that participants rated the evaluation session as more temporally demanding ($M = 10.24, SD = 4.19$) than the practice session ($M = 8.28, SD = 4.09$), $F_{(1,54)} = 12.79, p < 0.01, MSE = 9.07$. Ratings of the frustration scale also yield no significant effects for highlighting, phrase set, or session. It is interesting that there are no effects for session (either a main effect or an interaction with phrase set or highlighting). This result seems to suggest that when the help that was provided in the practice session (such as highlighting on or ordered phrase set) was removed, participants did not feel more discouraged or stressed in the evaluation session.

6.4 Comparison to Previous Work

Data from our previous study on Twiddler typing rates[9] can be used as a baseline against which to compare our current typing rates. Although many differences exist between the two studies which could account for differences in typing rates (e.g., compensation, instructions, error highlighting, phrase set, etc.) we believe the comparison can still be illuminating. In order to compare the two studies we utilized a 2 (session) x 2 (study) ANOVA. The study factor has two levels: previous and current, which correspond to the original study and the current study. This analysis combines the current study's experimental conditions into one group. There is a significant interaction between the session and study factors, $F_{(1,68)} = 27.51, p < 0.01, MSE = 1.19$. A simple effects analysis shows that within the practice session, the current study yielded faster typing rates ($M = 5.99, SD = 1.75$) than the previous study ($M = 4.27, SD = 1.35$), $F_{(1,68)} = 8.84, p < 0.01, MSE = 2.88$. However, within the evaluation session there is no significant difference in typing rates between the current study ($M = 6.15, SD = 1.89$) and the previous study ($M = 7.18, SD = 2.08$), $F_{(1,68)} = 2.54, p = 0.12, MSE = 3.63$. In the previous study, typing rates increased significantly from the practice session to the evaluation session, $F_{(1,68)} = 35.81, p < 0.01$. However, in the current study, typing rates did not significantly change between the two sessions, $F_{(1,68)} = 0.61, p = 0.44$. Together, these results suggests that the current study raised typing rates in the first 20 minutes.

In order to investigate the possiblity that our Twiddler phrase set (as opposed to the MacKenzie phrase set) is responsible for the difference in typing rates for the first

condition, we compare our baseline condition (highlighting off and unordered phrase set) to the previous study's data. If there is a difference between baseline conditions we can attribute the change to any of the several differences between the two studies, including the phrase set. We used the same 2 (session) x 2 (study) ANOVA analysis strategy, but limited our data set to the baseline condition in our current study and the previous data's study. We therefore used only the data from 10 participants in the new study and the data from the 10 participants in the old study. Like before, we found a significant interaction between study and session, $F_{(1,18)} = 5.32, p = 0.03, MSE = 4.87$. A simple effects analysis of study at each level of session shows the practice condition does not have a statistically significant difference between the the old study $(M = 4.27, SD = 1.35)$ and the new study $(M = 4.17, SD = 1.74)$, $F_{(1,18)} = 1.69, p = 0.21, MSE = 2.41$. Likewise, in the evaluation condition there is no reliable difference between the old study $(M = 7.18, SD = 2.08)$ and the new study $(M = 6.69, SD = 1.87)$, $F_{(1,18)} = 0.31, p = 0.59, MSE = 3.91$. This result suggests that the phrase set by itself was not enough to alter typing rates across studies.

7 Discussion

Across all of our measures, the effects of our two aids are encouraging. In general, using the ordered phrase set and highlighting helps novice Twiddler typists' performance. The ordered phrase set increases typing rates and lowers the subjective physical demand during the practice session. While this effect did not persist in the evaluation session, increasing performance while using the aid may help adoption. Simply presenting the keys to be learned sequentially, in groupings that correspond to the keyboard layout, allows individuals with no experience to type meaningful phrases faster and with less effort. This result is consistent with existing research that has found training beginners on parts of a task, rather than the whole task is beneficial.

Enabling highlighting for the first 20 minutes of typing increases typing rates, reduces the number of errors, and reduces subjective ratings of effort. However, we believe that the results indicate that this highlighting may have a slight cost. Error rates increased and typing rates decreased once highlighting was turned off. While error rates increased, the group with highlighting made no more errors than the groups without; although typing rates decrease, they are not slower than the other groups. Using highlighting with a delay did not seem to have an overall positive effect on typing or error rates. It might be that we did not have the correct delay timing to show any meaningful benefit. The failure to find any significant benefits should not rule out future investigations into the utility of delaying highlighting for novices.

Comparing the data from this study to the first two sessions of our previous Twiddler evaluation shows that our aids are beneficial for the first twenty minutes of typing and do not hinder the second twenty minutes once removed.

We believe these aids would be helpful in convincing prospective users that a Twiddler is easy to adopt even though one types differently than current mobile phones. For example, in a mobile phone store, a demonstration that featured highlighting might entice potential users to try typing with chords. Once the user bought a mobile phone, a typing tutor on the phone could use the reduced phrase set to provide the user with

a quick feeling of accomplishment. Then, as the user became more experienced, the MacKenzie phrase set could be used to further the user's skill.

8 Future Work

We are interested in exploring the long term effects of our aids on novice performance. We would like to determine if extending the duration of using our aids might continue to improve novice typing rates. We would also like to explore the potential impact multi–character chords (MCCs) might have on novice users [8]. MCCs can generate a sequence of characters such as "the " or "ing ". Using MCCs reduces the number of chords that need to be typed and therefore can enhance typing rates. It would be interesting to see if the benefit of enhanced typing rate outweighs the cognitive cost associated with learning extra chords. We are also interested in exploring more familiar designs that incorporate chording similar to that of the Twiddler. While improving novice performance might help adoption, the physical device itself is also important. We are exploring a mobile phone based on the current Twiddler keyboard (Figure 4). For messaging or learning to type, a high resolution screen could be used for a tutorial program similar to our evaluation software which incorporates our typing aids. Given the rapid ability to type, this device might enable advanced mobile phone features such as mobile email.

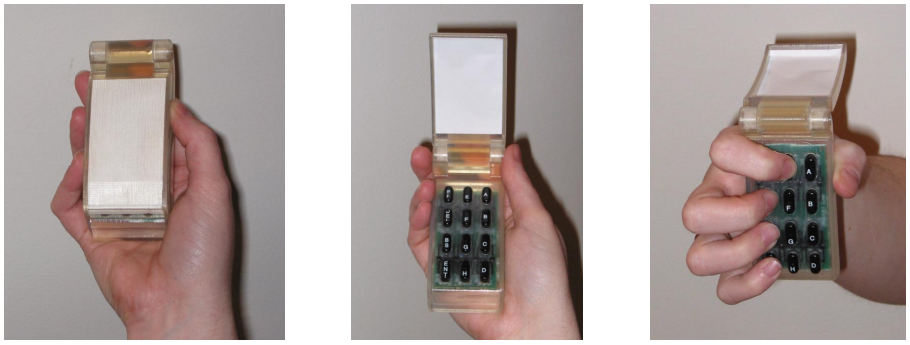

Fig. 4. A mobile phone design which incorporates chording capabilities.

9 Conclusion

In this paper, we presented a study examining two aids designed to help novice typists on the Twiddler mobile one–handed chording keyboard. We found that using an ordered phrase set designed around the Twiddler keymap helps typing rates and reduces physical demand. Using highlighting with our on–screen representation does hinder typing rates once turned off. However, having highlighting on reduces error rates and decreases the subjective physical demand. Given an expert Twiddler user's ability to enter text rapidly in a mobile setting and the ability to help novice typists with our two software aids, chording seems to be a viable mechanism for text entry on future mobile devices.

10 Acknowledgements

This work was funded in part by NSF Career Grant #0093291 and the NIDRR Wireless RERC and thanks to Stephen Griffin for his help with the phone prototype.

References

1. S. Baker, H. Green, B. Einhorn, M. Ihlwan, A. Reinhardt, J. Greene, and C. Edwards. Big bang! BusinessWeek, June 2004.
2. R. Catrambone and J. Carroll. Learning a word processing system with training wheels and guided exploration. In *Proceedings of CHI 1987*, pages 169–174. ACM Press, 1987.
3. S. G. Hart and L. E. Staveland. *Human mental workload*, chapter Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. North-Holland, 1988.
4. S. G. Hill, H. P. Iavecchia, J. C. Byers, A. C. Bittner, A. L. Zaklad, and R. E. Christ. Comparison of four subjective workload rating scales. *Human Factors*, 34(4):429–439, August 1992.
5. C. L. James and K. M. Reischel. Text input for mobile devices: comparing model prediction to actual performance. In *Proceedings of CHI 2001*, pages 365–371. ACM Press, 2001.
6. A. Kirlik, A. D. Fisk, N. Walker, and L. Rothrock. *Making decisions under stress: Implications for individual and team training*, chapter Feedback Augmentation and Part-Task Practice in Training Dynamic Decision-Making Skills. American Psychological Association, 1998.
7. M. Lindstom. Message madness our big chance. SMH http://www.smh.com.au, February 2002.
8. K. Lyons, D. Plaisted, and T. Starner. Expert chording text entry on the twiddler one–handed keyboard. In *Proceedings of ISWC 2004*, 2004.
9. K. Lyons, T. Starner, D. Plaisted, J. Fusia, A. Lyons, A. Drew, and E. Looney. Twiddler typing: One-handed chording text entry for mobile phones. In *Proceedings of CHI 2004*. ACM Press, 2004.
10. I. S. MacKenzie, H. Kober, D. Smith, T. Jones, and E. Skepner. Letterwise: prefix-based disambiguation for mobile text input. In *Proceedings of UIST 2001*, pages 111–120. ACM Press, 2001.
11. I. S. MacKenzie and R. W. Soukoreff. Phrase sets for evaluating text entry techniques. In *CHI '03 extended abstracts*, pages 754–755. ACM Press, 2003.
12. I. S. MacKenzie and S. X. Zhang. The design and evaluation of a high-performance soft keyboard. In *Proceedings of CHI 1999*, pages 25–31. ACM Press, 1999.
13. Mobile CommerceNet http://www.mobile.seitti.com, January 2002.
14. R. W. Soukoreff and I. S. MacKenzie. Metrics for text entry research: an evaluation of msd and kspc, and a new unified error metric. In *Proceedings of CHI 2003*, pages 113–120. ACM Press, 2003.
15. D. Wigdor and R. Balakrishnan. TiltText: Using tilt for text input to mobile phones. In *Proceedings of UIST 2003*. ACM Press, 2003.
16. D. Wigdor and R. Balakrishnan. A comparison of consecutive and concurrent input text entry techniques for mobile phones. In *Proceedings of CHI 2004*, pages 81–88. ACM Press, 2004.
17. Y. Yeh and C. D. Wickens. Dissociation of performance and subjective measures of workload. *Human Factors*, 30(1):111–120, February 1988.