AN EVALUATION OF TEXT-ENTRY IN PALM OS – GRAFFITI AND THE VIRTUAL KEYBOARD

Michael D. Fleetwood, Michael D. Byrne, Peter Centgraf,
Karin Q. Dudziak, Brian Lin, and Dmitryi Mogilev
Department of Psychology MS-25
Rice University
Houston, TX 77005 USA
fleet@rice.edu

The handheld organizer or personal digital assistant (PDA) is rapidly becoming a popular organizational tool, and there is a need for evaluation of alphanumeric character entry on these devices. The Palm operating system, the most common PDA operating system on the market, uses two methods for character entry, an on-screen virtual keyboard and a single-character handwriting recognition system called Graffiti. An initial experiment was conducted to investigate the character entry rates of novice and expert users of the device for the two methods of input. Experts were found to reach an average rate of 21 words per minute (wpm) using Graffiti and 18 wpm using the virtual keyboard. Novices were able to use Graffiti at a rate of 7 wpm and the virtual keyboard at 16 wpm. These character entry rates are evaluated with respect to some theoretical limitations, a predicted rate of entry based on Fitts' and the Hick-Hyman laws for the virtual keyboard, and pen and paper printing for Graffiti. The potential gain for new character entry systems and opportunities for improvement are discussed.

INTRODUCTION

The handheld organizer or personal digital assistant (PDA) is rapidly becoming a popular organizational tool, replacing traditional pen and paper methods in all age ranges. As with other types of newly developing portable devices, such as the mobile phone, the issue of text entry on these devices has become a prominent one. A number of different methods are currently available for text-entry on PDAs, with new ones being developed every day. Designers, researchers, and users would all like to gain some insight as to the relative efficiency of these different methods for text-entry. This set of studies was developed with that goal in mind.

Of the different operating systems currently available on PDAs, roughly 85 percent of handheld PDAs sold use the Palm operating system (Palm OS) from Palm Computing (Consumer Reports, 2001). Data can be entered on these units by tapping on an on-screen keyboard (referred to as a "virtual" or "soft" keyboard) or writing in a shorthand known as Graffiti, although other methods and different handwriting-recognition software are becoming more readily available.

The initial goal of this line of research is to provide an estimate of character entry rates using these two input methods, Graffiti and the virtual keyboard. Palm Computing suggests that a rate of 30 words per minute is possible (Palm Computing, 1995), and we'd like to begin to evaluate that claim. Beyond that, we hope to apply the data we gather in a broader scope. For one, we would like to use our measurements in future analyses of data-entry tasks. For example, a measure of the time to enter a character using either method may be used in a GOMS style analysis (Card, Moran & Newell, 1983) to predict performance in character input tasks in the PDA environment. An experiment in which participants were asked to enter characters and numbers into Palm OS handhelds was designed to provide such data. Second, we would like to use our evaluation of the methods as a benchmark for comparison with other devices and methods for text-entry on PDAs. Such a benchmark should provide us with information that will guide the development of future innovations, i.e. if we invest in developing new methods for text-entry, how much can we expect to improve over the existing ones? As a benchmark for Graffiti, the results of the first experiment were coupled with the data gathered in a second experiment, in which the time to print characters was measured. With respect to the virtual keyboard, the benchmark calculated is a predicted rate of entry, based on Fitts' law for rapid aimed movements and the Hick-Hyman law for choice selection time.

Considering that character input times and preferences are highly likely to be different for "experts" and "novices", Experiment 1 was structured around these two groups of users. Additionally, a careful error data analysis was performed to investigate the possible correlation between the number of input errors that the participants had committed and their respective level of expertise.

EXPERIMENT 1

Method

Participants. 48 people volunteered for the experiment. The users were separated into two groups, novices and experts. An expert was defined as anyone who had owned a Palm OS handheld for 3 or more months (this amounted to a distinction based on use of Graffiti, as each of the expert participants used Graffiti as their primary method of text entry). Justification for such a definition was provided by the data, as there was a clear break in performance using Graffiti between participants who had used a Palm OS handheld for 3+ months and those that had not used one before. We found that the times between people who had owned a Palm OS handheld for 3-6 months and those who had owned it for 6+ months was negligible. No users in the 1 to 3 month range were tested. All of the novices had never before used a Palm OS handheld device.

Materials. Each participant entered three phrases into PDAs running Palm OS 3.1 as the operating system. A stopwatch was used to capture input times. The key for inputting the Palm OS Graffiti alphabet was also provided.

An index card containing three test phrases was used, two of which were designed to be representative of the types of phrases users might enter into a PDA and one that contained all 26 letters of the alphabet. The three phrases used in the experiment were, "meet subject in lab", "quick brown fox jumped over the lazy dog", and "504 983 2761".

Procedure. The Palm OS handheld was set up so that the participant could begin entering characters immediately in a mode that had been determined randomly (either Graffiti or the virtual keyboard). All participants completed a practice trial before beginning the timed phrases. The practice trial consisted of entering the alphabet, A through Z, and the digits 0-9. This practice trial was not timed. Upon completion of the practice trial, the phrases were entered in the same order for each participant. The order of the phrases was "meet subject in lab," "quick brown fox jumped over the lazy dog," and then "504 983 2761." After entering all the phrases with both input methods the participant filled out a questionnaire regarding their demographic data and preferences for the two data-entry methods.

Error and Character Coding. The number of errors, backspaces, and total number of characters entered were recorded to enable a time per character and an error rate to be calculated. Errors were counted by comparing the correct phrase to what was actually entered. The number of errors were counted to provide the lowest count possible in a manner approximating the Levenshtein minimum string distance (Soukoreff & MacKenzie, 2001). Only errors of commission, errors in the entry of characters, were counted; errors of omission, such as when a participant forgot to enter a word, were not added to the error count (although they were considered in the entry rate calculations).

Results

As an aid to GOMS style analyses, the mean time per character in seconds and corresponding standard deviation are presented in Table 1. Mean errors collapsed across the three input phrases are also presented in Table 1.

Figure 1 shows the average words per minute entry rate as a function of input method and level of expertise. Figure 1 illustrates that experts are faster than novices using both input methods, as indicated by a main effect of level of expertise, F (1, 45) = 85.56, p < 0.001. An analysis of simple main effects using *t*-tests further confirmed that experts were faster using both Graffiti, *t*(45) = 9.06, p < 0.001, and the virtual keyboard, *t*(45) = 2.86, p < 0.01. The MANOVA also revealed a significant interaction between level of expertise and input method, F (1, 45) = 87.30, p < 0.001. This interaction suggests that experts were faster using Graffiti than the virtual keyboard, while novices were slower using Graffiti than the virtual keyboard.

Method	Expert	Novice
Graffiti (spc)	0.58 (0.11)	1.76 (0.55)
Virtual Keyboard (spc)	0.67 (0.12)	0.78 (0.14)
Graffiti (errors)	1.67 (1.33)	2.27 (2.21)
Virtual Keyboard (errors)	0.21 (0.26)	0.53 (0.48)

Table 1. Mean seconds per character (spc) for the two input methods, collapsed across the three phrases (with standard deviations in parenthesis). Mean number of errors (with standard deviations) collapsed across the three test phrases are also presented.

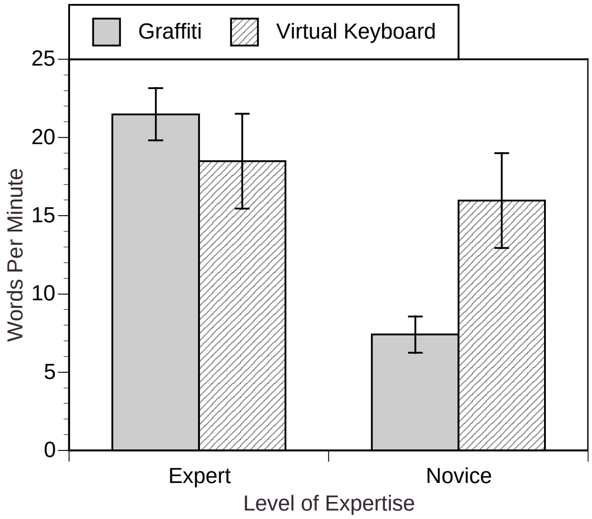

Figure 1. Overall mean words per minute rate by level of expertise and input method. Error bars represent the 95% confidence interval.

Errors were also analyzed as a function of level of expertise and input method. As indicated in Figure 2, both experts and novices made significantly more errors using Graffiti than using the virtual keyboard, F (1, 45) = 33.42, p < 0.001. Interestingly, the data did not reveal a reliable difference in the number of errors committed by experts and novices, F (1, 45) = 2.75, p = 0.11.

Although the effect is smaller in magnitude, an effect of test phrase was also revealed in the analyses, F (1, 45) = 4.56, p = 0.038, indicating that participants had a longer average time per character on the longer sentence than on the short sentence. This same effect did not quite reach our predetermined level of significance (p = 0.05) for the virtual keyboard, F (1, 45) = 3.94, p = 0.053.

An analysis of the participants' responses on the questionnaire regarding their subjective ratings for which method was more efficient and which they preferred was also conducted based on 2X2 frequency tables. The participants' subjective rating of the more efficient input

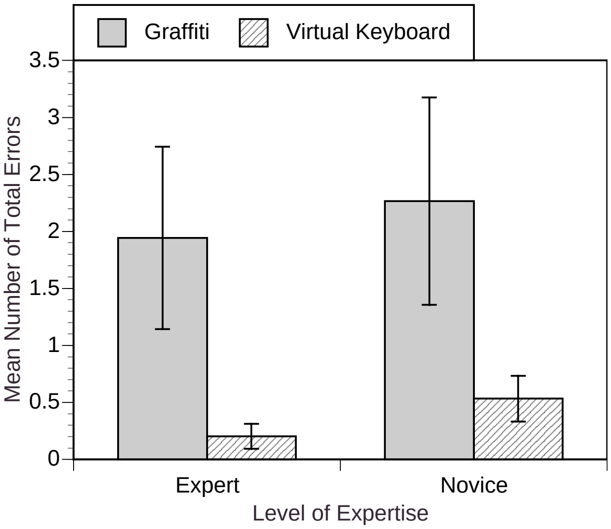

Figure 2. Mean number of total errors by level of expertise and input method. Error bars represent the 95% confidence interval.

method was reliably associated with their actual efficiency, $\chi^2(1) = 6.35$, $p = 0.01$. The input method preferred by participants was also reliably associated with both their subjective and objective efficiency, $\chi^2(1) = 14.82$, $p < 0.001$ and $\chi^2(1) = 13.94$, $p < 0.001$, respectively (see Table 2). The tendency for Experts to prefer Graffiti and Novices to prefer the on-screen virtual keyboard was significant at the .001 level, $\chi^2(1) = 15.63$ (Table 3).

	Faster using GR	Faster using VK
Preferred GR	18	12
Preferred VK	0	17

Table 2. Frequency data relating the number of participants who preferred a method of input (GR = Graffiti and VK = virtual keyboard) and the number of participants who inputted text quicker with a particular method, illustrating that participants generally preferred the method that they were fastest with.

	Expert	Novice
Preferred GR	21	9
Preferred VK	1	16

Table 3. Frequency data relating the number of participants who preferred a method of input (GR = Graffiti and VK = virtual keyboard) and the level of expertise of the participants, illustrating that experts generally preferred Graffiti and novices, although more split on the issue, generally preferred the virtual keyboard.

Discussion of Character Entry Rates

Our study suggests that experience is a key factor in predicting text input rates using a Palm OS handheld device. As expected, Graffiti rates of entry (WPM) appear to increase dramatically with prolonged use, whereas virtual keyboard rates remain relatively flat regardless of user experience. This could be expected because the "experts" in the study nearly all used Graffiti as their primary method of text entry. It is also possible that virtual keyboard times are limited primarily by the physiological limitations of finding and selecting targets as expressed by Fitts' Law, while Graffiti times are initially limited by lack of experience with the unique character system. Over time, users appear to acclimate to the new letterforms and are able to recall and create them more quickly. Novice users are dramatically faster when using the simpler virtual keyboard method, but users familiar with Graffiti are able to input text somewhat more rapidly than with the virtual keyboard. Graffiti seems to have a non-trivial learning curve but can be faster for users who make the effort to learn.

The observation that experts were found to be faster than novices on the virtual keyboard is an intriguing one. Clearly, our distinction between the expert and novice categories is based on the participants' use of Graffiti, and all users indicated that they were familiar with the QWERTY keyboard. We could speculate then on the reasons for Graffiti experts outperforming Graffiti novices on the virtual keyboard (confidence and/or comfort with the device, practice using a stylus, etc.), but more research is needed to bear out the cause of this discrepancy in performance between the two groups.

Analysis of error data shows a more direct contrast between input methods. Graffiti input shows a significantly higher rate of errors (9%) than virtual keyboard input (2%) for both experts and novices. It appears that while Graffiti users can gain speed with practice, they aren't able to increase their accuracy. This is consistent with other text-entry experiments that have found that subjects did not improve their accuracy with practice, but did get faster at the task (MacKenzie, Nonnecke, McQeen, Riddersma & Metz, 1994). Of course, the type of errors made by experts and novices may be qualitatively different; perhaps experts trade speed for accuracy while novices are simply less proficient with a stylus.

Regarding user preference, users were fairly accurate in identifying which input method was most efficient for their own use, and not surprisingly, they tended to prefer the faster method. It is useful to note that those users who preferred the method in which they were slower were all novice users who enjoyed using Graffiti. This suggests that many novices will use Graffiti because of the novelty, despite the initial learning curve, and will subsequently become more efficient with that method.

EVALUATION OF GRAFFITI

One of our initial goals for this program of research was to gain some insight as to the effectiveness of Graffiti and the virtual keyboard relative to other methods of data entry. One question to ask in this realm might be how much can Graffiti be improved upon if further design iterations are carried out. Ideally, we would investigate this issue by comparing the effectiveness of the two input methods to some "best case scenario" or theoretical upper bound limit of performance. It is difficult to define exactly what the "best" method of text entry is, since new systems of character recognition software are continually being developed. Fortunately, we can compare it to a method that is currently widely used and meets the same restrictions as those imposed on Graffiti, where each character is entered individually and can be written independently of other characters: printing Roman letters with pen and paper.

Graffiti already capitalizes on prior learning of printing in English (MacKenzie & Zhang, 1997), which comprises a major advantage of using Roman letters as a basis for a character recognition system. However, there are several reasons why normal print does not make a good candidate for a character recognition system (Goldberg & Richardson, 1993). For one, there are a number of characters that require multiple strokes, making it difficult for a character recognition system to determine where one character ends and another begins. Also, print characters are not well separated in "sloppiness space," i.e. they are not robustly distinguishable when written sloppily. On the other hand, unistroke systems of character entry, such as Graffiti, have been designed to take advantage of prior learning of print while minimizing the aforementioned disadvantages of using English printing as an input method on a PDA. Additionally, there are several unique advantages to using a unistroke character entry system, such as efficient use of screen real estate and "eyes free" operation (Goldberg & Richardson, 1993). Graffiti has also been revered as a theoretically faster method of text entry than print (MacKenzie & Zhang, 1997).

Despite the fact that using printing as a benchmark for our evaluation comes with several caveats, there are reasons why printing provides a logical benchmark. Foremost, it requires no additional learning and it is widely used. Additionally, Goldberg (Goldberg & Richardson, 1993) discussed a trade-off between character entry speed and ease of learning. By capitalizing on previous learning of printing in English, i.e. because the characters were designed to mimic Roman letters as closely as possible (MacKenzie & Zhang, 1997), Graffiti seems to lie towards the ease of learning end of this spectrum. (In contrast, phonetic-based systems, such as many secretarial shorthand systems, can achieve much higher entry speeds, but at the cost of learning time.) Because Graffiti is based on the Roman alphabet, printing is already well associated with Graffiti and provides one logical option as a benchmark of the new system's effectiveness.

Also, there are a number of studies that have measured the printing speeds of native English speakers, giving us a good starting point for our evaluation. The studies have specified a relatively wide range of printing rates, from 13 to 22 words per minute (Card, Moran & Newell, 1983). In order to give us a direct comparison with Graffiti as it was evaluated in Experiment 1, a second experiment was conducted.

Design

Participants were simply asked to print each phrase written on the index card in the same order as in Experiment 1. The time to print each phrase was recorded using a stopwatch.

Results

Character entry rates for printing were compared with entry rates for experts using Graffiti and the same Graffiti experts using the virtual keyboard. The average wpm entry rate for pen and paper printing observed in the study was 26.8 wpm with a standard deviation of 3.8 wpm, which shown in Figure 3 with the data from the first experiment.

A between-subjects ANOVA was calculated to investigate performance across the three levels of input method (Graffiti, virtual keyboard, and pen and paper printing). (Note: Graffiti character entry rates and virtual keyboard entry rates are derived from the same group of participants (expert Graffiti users) and could have been analyzed using a within-subjects analysis as in the first experiment. The between-subjects ANOVA used here is a more conservative approach that allowed for easy incorporation of a separate group, print handwriting.) A significant effect of input method was revealed, F (2, 65) = 28.58, $p < 0.001$, indicating that participants differed in their average character input time across the three input methods. A *t*-test was also conducted to examine the differences between Graffiti and print. This confirmed that subjects were faster when entering data using print than using Graffiti, t(44) = 4.74, $p < 0.01$.

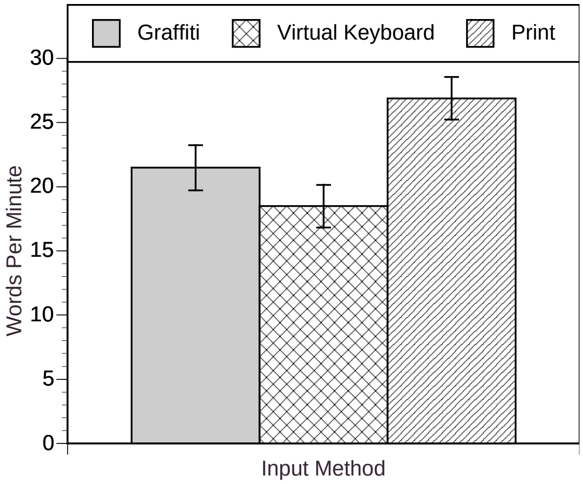

Figure 3. Mean words per minute entry rate for pen and paper printing, Graffiti, and the virtual keyboard. Error bars represent the 95% confidence interval.

EVALUATION OF THE VIRTUAL KEYBOARD

Virtual keyboards have been examined in a number of previous studies (MacKenzie & Zhang, 1999; MacKenzie, Zhang & Soukoreff, 1999; Soukoreff & MacKenzie, 1995). However, these evaluations have not examined virtual keyboards on the scale of PDAs. One of the methods of examining soft keyboards is through a quantitative analysis based on Fitts' law for physical movements to a target and the Hick-Hyman law for choice selection time (Soukoreff & MacKenzie, 1995). Using these quantitative formulas as a basis for examination we calculated theoretical upper bound and lower bound limits for performance on a virtual keyboard.

Our upper bound prediction was calculated using the following equation based on Fitts' law for the movement time (Mt) between any two keys (i and j):

$$Mt_{ij} = 0.204 \log_2 ((A_{ij}/W_{ij}) + 1)$$

where A is the distance between keys, measured on the screen of the PDAs in pixels, and W is the size or width of the target key, also measured in pixels.

The parameter in the equation, Fitts' law slope (0.204), is based on a study of the bandwidth for pointing tasks using a stylus as a computer input device (MacKenzie et al. 1991), which found 4.9 bits per second (bps) to be an appropriate value for tasks of this nature (1/4.9 = 0.204). In the one instance where a key is selected twice in sequence (the "e"s in "Meet") (i.e. where there is no movement to a new target) 0.153 seconds is used as the MT. This is the value estimated by Soukoreff and

MacKenzie (1995) in their study of virtual keyboards and approximates the value of 140 ms used by Card et al. (1983, 60) for a typist repetitively pushing a key with a finger.

Our calculated theoretical upper bound for the Palm OS virtual keyboard is 30.2 wpm for the short phrase and 27.3 wpm for the long phrase. This theoretical maximum rate of entry represents the time to physically input the phrase assuming no time for visual search or decision making. To calculate a lower bound prediction we add in a parameter for decision making and visual search based on the Hick-Hyman equation for choice reaction time, which represents the predicted time for novices to visually scan a 27-key layout to find a target key (calculated as 0.951 seconds). The lower bounds were calculated as 8.9 wpm and 8.6 wpm for the short and long phrases respectively. The performance of users in our experiment on the virtual keyboard, at 16 wpm for Graffiti novices and 18 wpm for Graffiti experts, falls well within these theoretical bounds.

Several aspects of these calculations are worth noting. For one, the difference between the calculated bounds for the two phrases indicates that the distances between the letters are longer on average for the long phrase than for the short phrase. In this sense, the long phrase is more difficult to input than the short phrase, and provides some explanation as to why we observed lower wpm rates for the long phrase than the short phrase on the virtual keyboard.

Several other studies have examined user performance on a virtual keyboard. The character entry rates observed in these studies fell at 22.9 wpm (Mackenzie, et al., 1994) and 20.2 wpm (MacKenzie & Zhang, 1999). The latter study is probably most comparable to the results of this study, as it represents user performance on a "quick test" where users were not given substantial practice using the input method before testing. The rate observed here, 18 wpm, is only slightly lower (possibly due to the layout of the keyboard and the additional "keys" on the Palm OS keyboard) but qualitatively seems to lie within the same range.

Discussion of Evaluations

As mentioned previously, there are several reasons in favor of and against considering print as a benchmark for a single character entry system on PDAs. If we do use it as our benchmark then the question becomes how far our current methods of character entry, Graffiti in this case, are off from this goal. Stated another way, we might ask how much do we potentially have to gain if we redesign our current input methods or develop new ones. Based on the results of these experiments, there is about 5 wpm separating expert Graffiti users and print, from about 21 to 26 wpm. Other experiments have put print speeds at a lower rate, 13 to 22 words per minute (Card, Moran & Newell, 1983), in which case the gap separating Graffiti and print decreases substantially (or the relationship even reverses).

These findings have a couple of implications for designers. They provide some information as to whether or not it is worth the effort to improve upon Graffiti as a single character input method for PDAs. Also, they give us some idea as to where improvement can be made. For any real increase in character entry speed, we may have to move away from orthographic single character entry systems, as there may not be much room for improvement here. Other options might include a phonographic system, such as a form of secretarial shorthand, or a comprehensive handwriting recognition system for cursive handwriting. However, as previously discussed, there are several aspects of a single character entry system that make it well-adapted for use on PDAs (limited screen real estate, "sloppiness" space, ease-of-learning), which may act as barriers too the development of new entry methods. Of course, these same aspects act as bounds on the system and limit its potential. Given these bounds, Graffiti seems to approach the upper limits of character entry rate that can be achieved with such a system. (There may be more room for improvement on other aspects of the system, such as user comfort, user preference, and rate of learning.)

The limitations that constrain performance on the virtual keyboard are relatively well defined, specifically Fitts' law for rapid aimed movements and the Hick-Hyman law for choice selection time. Indeed, performance in our study fell well within the predicted range based on these laws. As noted by other studies (Soukoreff & MacKenzie, 1995), such predictability can prove quite useful in the evaluation of different types of soft keyboards. Our calculation of a theoretical upper bound performance level of 27 wpm suggests that user performance could improve with practice. This theoretical upper bound also suggests that a virtual keyboard "expert" could theoretically outperform a Graffiti expert in character entry rate, and likely do it with much fewer errors. Of course, the question of whether this upper bound limit can actually be reached in a reasonable course of time needs further research.

Gains in character entry speed also can be made by designers through a reorganization of the virtual keyboard itself. Indeed, several researchers have worked towards the development of a more efficient soft keyboard (MacKenzie & Zhang, 1999; Zhai & Barton, 2001), and have provided evidence that an optimized layout can substantially improve performance.

REFERENCES

Card, S. K., Moran, T. P., and Newell, A. (1983). *The Psychology of Human-Computer Interaction*. (Lawrence Erlbaum, Hillsdale, NJ).

Consumer Reports. (May 2001). Data to go. *Consumer Reports*, 66(5), 20-23.

Goldberg, D. and Richardson, C. (1993). Touch-typing with a stylus. Proceedings of the INTERCHI '93 Conference on Human Factors in Computer Systems, 80-87.

MacKenzie, I. S., Nonnecke, R. B., McQueen, J.C., Riddersma, S., & Metz, M. (1994). A comparison of three methods of character entry on pen-based computers. *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting*, 330-334.

Mackenzie, I. S. Zhang, S. X.. (1997). The immediate usability of Graffiti. *Proceedings of Graphics Interface '97*, 129-137.

Mackenzie, I. S. Zhang, S. X.. (1999). The design and evaluation of a high-performance soft keyboard. *Proceedings of the SIG-CHI Conference on Human factors in computing systems,* 25-31.

Mackenzie, I. S. Zhang, S. X. and Soukoreff, R. W.,(1999). Text entry using soft keyboards. *Behaviour & Information Technology*, 18(4), 235-244.

Palm Computing. (1995, January). Suddenly Newton understands everything you write. Pen Computing Magazine, p. 9.

Soukoreff, R. W. and Mackenzie I. S. (1995). Theoretical upper and lower bounds on typing speed using a stylus and a soft keyboard. *Behaviour & Information Technology*, 14(6), 370-379.

Soukoreff, R. W. & MacKenzie, I. S. (2001). Measuring errors in text entry tasks: An application of the Levenshtein String Distance Statistic. *Extended Abstracts of the SIG-CHI Conference on Human factors in computing systems,* 319-320.

Zhai, S. & Barton, S. A. (2001). Alphabetically biased virtual keyboards are easier to use layout does matter. *Extended Abstracts of the SIG-CHI Conference on Human factors in computing systems,* 321-32.