Using Paper Mockups for Evaluating Soft Keyboard Layouts

I. Scott MacKenzie

Unit for Computer-Human Interaction (TAUCHI)
Dept. of Computer & Information Sciences
FIN-33014 University of Tampere
Tampere, Finland
+358 3 215 8566

Dept of Computer Science
York University
Toronto, Ontario, Canada M2J 1P3
+1 416 736 2100
smackenzie@acm.org

ABSTRACT

Two soft keyboard layouts were evaluated for entry speed using paper mockups and hand timing. Twelve participants used a stylus to tap the well-known "quick brown fox" phrase five times on Qwerty and Opti layouts. Entry speeds differed significantly between the layouts: 29.5 wpm for Qwerty and 12.3 wpm for Opti. There was a significant improvement with practice over the five trails, particularly for the Opti layout (62.3%). The improvement was less with the Qwerty layout (35.3%), likely due to participants' prior experience with the layout. The merits and limitations of the evaluation method are discussed.

Keywords

Soft keyboards, text entry, evaluation methods, walk-up usability

INTRODUCTION

Developing efficient methods of text entry is a popular research topic in today's race for new mobile communications products. Given the ever-shortening time to market of new initiatives, developing efficient methods of evaluation is a desirable adjunct to research. This paper is primarily concerned with the latter of these two themes – to formally develop, test, and critique a rapid evaluation method for new input techniques. The problem is presented in the context of the former theme – the development of efficient means of text entry for mobile systems.

Mobile Text Entry

Despite the appeal of miniaturization, mobility bears a price. The physical means for input are constrained by the small form factor, and, so, desktop devices (full-size keyboards and mice) are not practical. Other input mechanisms are required, such as speech, physical keyboards with fewer or smaller keys, or stylus input.

Stylus-based mobile systems typically support two forms of input: gesture recognition and tapping. Stylus tapping on a graphic representation of a keyboard – a soft keyboard – is popular for text entry and is supported on all stylus-based mobile devices. A soft keyboard is easy to implement and provides an alternative to handwriting.

With physical keyboards, non-Qwerty layouts are of little interest today. Although alternate layouts, such as Dvorak [9], alphabetic [10, 18], or chord keyboards [2, 6], can support higher entry rates, substantial practice is required to gain proficiency. This, combined with a large installed base for Qwerty, has ensured the continued role of Qwerty as the keyboard of choice for desktop computing.

For soft keyboards, the arguments for Qwerty are diminished. Since the device is virtual rather than physical, manufacturing costs lie in the software, and, are one-time only. Thus, exploring the design space of soft keyboard layouts has emerged as a significant area of research [3, 4, 7, 8, 13, 15, 16, 20, 22-24].

Expert vs. Novice Users

Most work on the design of text input methods, such as soft keyboard layouts, focuses on the potential or expert entry rate of a design [3, 8, 13, 23]. However, the novice experience is paramount for the success of new text input methods [12, 14]. This is at least partially due to the target market. Mobile devices, such as mobile phones and PDAs, once specialized tools for professionals, are increasingly used by consumers. It follows that immediate or walk-up usability is important. In other words, it is a moot point to establish the expert text entry rate if prolonged practice is required to achieve it. Consumers, discouraged by their initial experience and frustration, may never invest the required effort to become experts.

Evaluation

Empirical evaluations of new interaction techniques are time consuming and labour-intensive. And so, a related research topic is the development of efficient methods of evaluation. There are a variety of such methods in use, such as "wizard of oz", where the user unwittingly interacts with a human instead of a system [1, 5]. Clearly this is efficient, since implementation is delayed until evidence is gathered on problems in the interface.

Paper mockups provide a convenient and efficient means to gather feedback from users. In this case, an interface is implemented on paper and user impressions are solicited, perhaps across several hypothetical implementations [17, 19]. Generally, such evaluations are qualitative, as performance measurements are difficult to gather.

Our interest here is to formally test the use of paper mockups in evaluating soft keyboard layouts. We are interested in a quantitative evaluation, since the most common research questions on soft keyboards pertain to text entry speed.

To increase the efficiency of the evaluation, our method involves the simultaneous testing of all participants, and engages the participants as experimental assistants in the evaluation. The method, results, and analyses are presented in the rest of this paper.

METHOD

Participants

Twelve volunteer participants (8 female, 4 male) were recruited from the local university. All were senior undergraduate or graduate students enrolled in the author's course in human-computer interaction. The participants also served as assistants in conducting the experiment (see Procedure below).

Apparatus

The Qwerty and Opti soft keyboards layouts were selected for evaluation. Opti is a high performance layout [13] designed using the Fitts-digraph model of Soukoreff and MacKenzie [20]. The predicted expert entry rates are 30 wpm for Qwerty and 42 wpm for Opti [11]. Both layouts were implemented as paper mockups. See Figure 1.

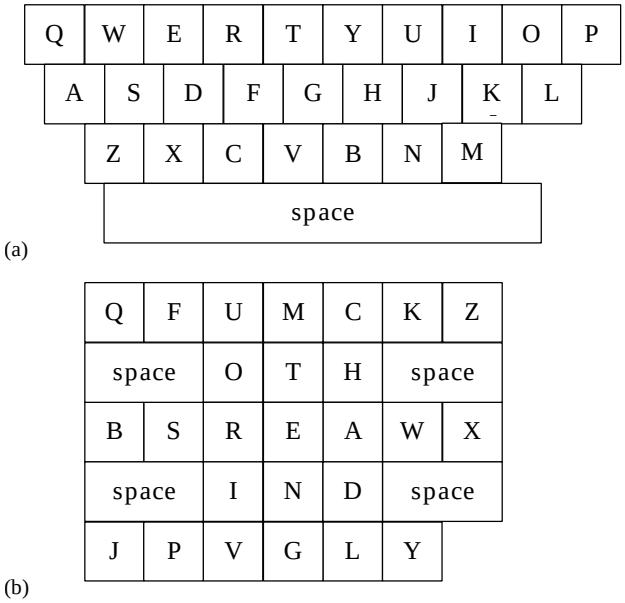

Figure 1. Soft keyboard layouts. (a) Qwerty (b) Opti

As measured on the paper mockup, the Qwerty layout was 9.1 x 3.6 cm, the Opti layout 7 x 5 cm. These dimensions are larger than typical for soft keyboards on PDAs. However, this should not impact performance as there is both theoretical and empirical evidence [14] that text entry rates for soft keyboards are not affected by the size of the layout.

With this highly-simplified apparatus, entry times could not be electronically measured, as there was no sensing technology or experimental software. Entry times were hand recorded with a timing device, such as a sports watch or mobile phone in stop watch mode.

Procedure

Participants were instructed to study and memorize the following 43-character phrase:[1]

`the quick brown fox jumps over the lazy dog`

The phrase was entered by tapping on the soft keyboard layout with a stylus. Participants provided their own stylus. Most used either a pen with the tip covered (or held upside down), or a pencil with the lead retracted. Used in this manner, the layout sheet remained clear of marks throughout the testing.

The instructions were to enter the phrase "as quickly as possible while trying not to make mistakes". Since no text was generated and accuracy was not recorded, some additional clarification was given on the need to proceed quickly (not recklessly) while accurately tapping the correct letters on the soft keyboard.

The participants worked in groups of two: one tapped while the other timed. A trial began when the timer said "start". Since no text was generated electronically, it was difficult for the timer to follow the progression of stylus taps. And so, participants were instructed to say "stop" upon tapping the last character (the "g" in "dog"). Timing was thus terminated for the phrase. The measurement in seconds was entered in a log sheet.

The procedure above was repeated five times using one layout, then five times using the other layout. Following this, the participants reversed their tapping and timing rolls and repeated the procedure.

To compensate for potential learning effects due to the order of testing layouts, participants were divided into two groups. Six participants entered with the Qwerty layout first, followed by Opti. The other half reversed the order.

The experiment was conducted in a classroom as part of a regularly scheduled lecture for a course in human-computer interaction. The total time to conduct the experiment was about 20 minutes.

Design

The experiment was treated as 2 ? 2 ? 5 mixed design. Group was a between-subjects factor with two levels (Group 1 vs. Group 2, six participants per group). The

[1] The following 45-character variant is sometimes used: "the quick brown fox jumped over the lazy dogs" [20]. In either case, the distinguishing feature is that the phrase contains every letter of the English alphabet. This ensures that every alphabetic key on the layout is tapped at least once. In fact, this phrase is somewhat atypical of English, since highly infrequent letters, such as "z", "x", and "q", are over-represented.

within-subject factors were Layout with two levels (Qwerty vs. Opti) and Trial with five levels (1, 2, 3, 4, 5). The total amount of input was 6 participants/group ? 2 groups ? 2 layouts ? 5 trials = 120 phrases.

Entry time was the only measurement taken. For each phrase, the entry time was converted to entry speed using (43 / 5) / (*t* / 60), where 43 is the size of the phrase in characters, 5 is the number of characters per word, *t* is the recorded entry time in seconds, and 60 is the number of seconds in a minute.[2]

RESULTS AND DISCUSSION

Counterbalancing the order of testing layouts achieved the desired result as the main effect and interactions for Group were not statistically significant.

The grand mean for entry speed was 19.4 wpm. The speed for the Qwerty layout was quite fast at 26.5 wpm, while that for the Opti layout was only 12.3 wpm. The difference was statistically significant ($F_{1,10}$ = 797.0, p < .0001).

There was considerable variation by participant. For Qwerty, participant means over the five trials ranged from 18.7 wpm to 30.2 wpm. For Opti, the means ranged from 6.7 wpm to 16.0 wpm. This suggests that participants approached the task with different attitudes on balancing speed with accuracy. The highest speeds recorded for single phrases were 35.0 wpm for Qwerty and 20.7 wpm for Opti.

There was also a significant effect for Trial ($F_{4,40}$ = 50.7, p < .0001), implying that participants' entry speed increased with practice. The Layout by Trial interaction effect was also significant ($F_{4,40}$ = 2.7, p < .05), although much less so than either main effect. The trends by Layout and Trial are seen in Figure 2.

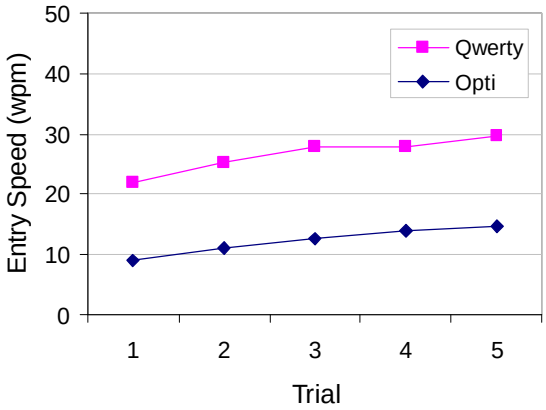

Figure 2. Entry speed (wpm) by Layout and Trial

[2] It has been a convention since about 1905 to standardize the computation of entry speed in "words per minute", where a word is defined as five keystrokes [21, p. 182]. This includes letters, spaces, punctuation, and so on.

Is the Comparison Fair?

The improvement with practice evident in Figure 2 illustrates an interesting phenomenon in evaluating soft keyboards. The relevant figures are summarized in Table 1.

Table 1
Improvement With Practice

Layout	Entry Speed (wpm)		Improvement
	Trial 1	Trial 5	
Qwerty	21.8	29.6	35.3%
Opti	9.0	14.6	62.3%

From Trial 1 to Trial 5, the participants' entry speed improved by 35.3% for the Qwerty layout. This is quite substantial, however, the improvement was even greater for the Opti layout, where entry speed increased by 62.3% from Trial 1 to Trial 5.

The relatively large improvements overall are likely due to participants becoming comfortable with the procedure (no practice trials were administered). However, the lower improvement for the Qwerty layout is likely due to participants' substantial prior experience with the layout. This experience means less visual scan time is required to "find the next letter". In essence, participants entered the experiment at different points on the learning curve for each layout. For Qwerty there is less "room for improvement", and, so, the percent improvement was much less.

Given the above, it is worth asking: Is the comparison fair? Is the comparison, as oft stated, "apples with apples"? On the one hand it is not, because participants' expertise is substantially different across the two layouts tested. (Because of their daily exposure to Qwerty, we sometimes affably accuse participants of "cheating for ten years" before showing up for the test!) On the other hand, the comparison is fair, since the procedure, in practical terms, accurately reflects the challenge facing designers of new soft keyboard layouts. Despite compelling evidence that alternative layouts can indeed promote higher entry speeds than Qwerty [11, 13, 23], the benefits only surface after substantial practice. This is a serious obstacle to acceptance, particularly in mobile computing where immediate or walk-up usability is important.

Is the Comparison Valid?

This research is motivated to demonstrate a simple method for evaluating soft keyboard layouts. While the goal of simplicity is clearly met, the method is only useful if the results have merit – if the results bear scrutiny in comparison with those obtained using a more realistic apparatus and a more thorough procedure. The layouts tested herein were chosen specifically to facilitate such a comparison. MacKenzie and Zhang [13] compared the Qwerty and Opti soft keyboard layouts in a longitudinal experiment using custom experimental software and a

Wacom tablet and stylus. The phrases of text presented to participants to enter were selected randomly from a set of 70 phrases. The average phrase length was 25 characters.

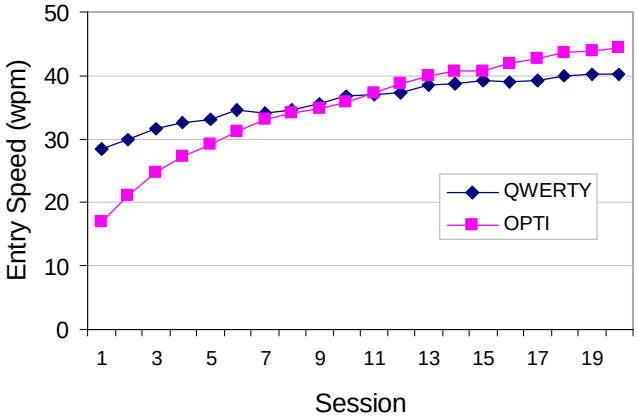

Figure 3. Qwerty vs. Opti results for entry speed from MacKenzie and Zhang [13]

The comparison of relevance here is with the session one of MacKenzie and Zhang's results. The important statistics are summarized in Table 2.

Table 2
Comparison Between Current Study and Session One Results from MacKenzie and Zhang [13]

Study	Phrases	Testing Time (minutes)	Entry Speed (wpm)	
			Qwerty	Opti
MacKenzie & Zhang	50-60	20-22	28	17
Current	5	3-5	26.5 (21.8 to 29.6)	12.3 (9.0 to 14.6)

While the results for the Qwerty layout are quite close between the two studies, the results for Opti are on the low side: 12.3 wpm in the current study compared with 17 wpm in MacKenzie and Zhang's study. For both layouts, further improvement with practice seems likely, given the trends in Figure 2. So, higher figures seem reasonable for the current study, had testing continued for 20-22 minutes, as in MacKenzie and Zhang's study.

Since the rate of improvement in the current study was greater with Opti (for reasons noted earlier), the mean would likely have been proportionally higher than for the Qwerty layout, perhaps settling in at the 17 wpm figure reported by MacKenzie and Zhang. It seems reasonable to conclude, therefore, that the results observed in the current study are consistent with those reported by MacKenzie and Zhang. In fact, if the goal is to measure walk-up entry rates, limiting the test to 3-5 minutes of input is arguably preferable. Results so gathered may be more representative of "walk-up" use those obtained over 20-22 minutes of testing.

Critiquing the Method

While the empirical results are reasonable (see above), they are quite limited since only one dependent variable was used. The method is clearly a compromise, and is not presented here as a substitute for a full and proper empirical study.

On the plus side, the experiment design and implementation were straight-forward, since a physical computing device was not used, nor was any software written. Furthermore, the procedure took only about 20 minutes, since participants were gathered together in a lecture hall and were tested together. There was a bit of confusion and noise while the mockup sheets were distributed and the procedure explained, but this can be corrected with careful advance planning. It would be useful, for example, for the experimenter to enlist an assistant to distribute the mockup sheets. Having the participants serve as assistants in gathering measurements seemed to work quite well. Once the experiment was underway, participants seemed to proceed without distractions.

It is extremely important that the experimenter create the correct atmosphere for the experiment. The present experiment seemed to succeed in this, however, this is partly because there were a small number of participants (12) and all were rather senior in their studies. The method might not work as well using, for example, a large class of first- or second-year undergraduate students.

Measuring Accuracy

While accuracy was not measured in this experiment, it may be possible to modify the procedure to capture errors. If participants use a felt-tip pen, their taps will leave a mark on the paper. Following the input of a phrase, the paper could be inspected for errors. The process would be tedious, and perhaps error prone in itself. Additionally, such a method requires a fresh keyboard rendering for each trial. However, this modification is something to consider for future use of the method. One potential benefit is that participants who are inclined to proceed recklessly, might be more careful if a record of their performance is generated.

CONCLUSIONS

We have demonstrated the use of paper mockups and hand timing to test soft keyboard layouts. Twelve participants were simultaneously tested, and also served as assistants in conducting the experiment. We measured 26.5 wpm for a Qwerty layout and 12.3 for an Opti layout. The results are reasonably consistent with those in a more formal empirical evaluation, suggesting that the methodology is useful as a quick and efficient means to test soft keyboard layouts.

ACKNOWLEDGEMENT

This research is sponsored by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the Academy of Finland (project 53796). This support is greatly appreciated. Thanks is also extended to the

students in the author's course "Research in advanced user interfaces: Models, measures, and methods" for participating in the experiment.

REFERENCES

1. Goldstein, M., Bretan, I., Sallnas, E.-L., and Bjork, H. Navigational abilities in audial voice-controlled dialogue structures, *Behaviour & Information Technology 18* (1999), 83-95.
2. Gopher, D., and Raij, D. Typing with a two-hand chord keyboard: Will the QWERTY become obsolete?, *IEEE Transactions of Systems, Man, and Cybernetics 18* (1988), 601-609.
3. Hughes, D., Warren, J., and Buyukkokten, O. Emprical bi-action tables: A tool for the evaluation and optimization of text input systems: Application I: Stylus keyboards, *Human-Computer Interaction 17* (2002), 271-309.
4. Hunter, M., Zhai, S., and Smith, B. A. Physics-based graphical keyboard design, *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems - CHI 2000*. New York: ACM, 2000, 157-158.
5. Klemmer, S. R., Sinha, A. K., Chen, J., Landay, J. A., Aboobaker, N., and Wang, A. Suede: A wizard of oz prototyping tool for speech user interfaces, *Proceedings of the ACM Symposium on User Interface Software and Technology -- UIST 2000*, New York: ACM, 2000) 1-10.
6. Kroemer, K. H. E. Operation of ternary chorded keys, *International Journal of Human-Computer Interaction 5* (1993), 267-288.
7. Lewis, J. R., Kennedy, P. J., and LaLomia, M. J. Development of a digram-based typing key layout for single-finger/stylus input, *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting*. Santa Monica, CA: HFES, 1999, 415-419.
8. Lewis, J. R., LaLomia, M. J., and Kennedy, P. J. Evaluation of typing key layouts for stylus input, *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting*. Santa Monica, CA: HFES, 1999, 420-424.
9. Lewis, J. R., Potosnak, K. M., and Magyar, R. L. Keys and keyboards, *Handbook on human-computer interaction*, ed. M. Helandar, T. K. Landauer, and P. V. Prabhu. Amsterdam: Elsevier, 1997).
10. MacKenzie, I. S., Nonnecke, R. B., Riddersma, S., McQueen, C., and Meltz, M. Alphanumeric entry on pen-based computers, *International Journal of Human-Computer Studies 41* (1994), 775-792.
11. MacKenzie, I. S., and Soukoreff, R. W. Text entry for mobile computing: Models and methods, theory and practice, *Human-Computer Interaction* (2002), [to appear].
12. MacKenzie, I. S., and Zhang, S. X. The immediate usability of Graffiti, *Proceedings of Graphics Interface '97*. Toronto: Canadian Information Processing Society, 1997, 120-137.
13. MacKenzie, I. S., and Zhang, S. X. The design and evaluation of a high-performance soft keyboard, *Proceedings of the ACM Conference on Human Factors in Computing Systems - CHI '99*. New York: ACM, 1999, 25-31.
14. MacKenzie, I. S., and Zhang, S. X. An empirical investigation of the novice experience with soft keyboards, *Behaviour & Information Technology 20* (2001), 411-418.
15. MacKenzie, I. S., Zhang, S. X., and Soukoreff, R. W. Text entry using soft keyboards, *Behaviour & Information Technology 18* (1999), 235-244.
16. MacKenzie, I. S., Zhang, X. I., and Soukoreff, R. W. Stylus tapping on a soft keyboard, *Behaviour & Information Technology 18* (1998), 235-244.
17. Nielsen, J. Paper versus computer implementations as mockup scenarios for heuristic evaluation, *Proceedings of IFIP INTERACT '90: Human-Computer Interaction*, 1990) 315-320.
18. Norman, D. A., and Fisher, E. Why alphabetic keyboards are not easy to use: Keyboard layout doesn't much matter, *Human Factors 24* (1982), 509-519.
19. Slaughter, L., Oard, D. W., Warnick, V. L., Harding, J. L., and Wilkerson, G. J. A graphical interface for speech-based retrieval, *Proceedings of the 3rd ACM International Conference on Digital Libraries -- DL '98*, New York: ACM, 1998) 305-306.
20. Soukoreff, W., and MacKenzie, I. S. Theoretical upper and lower bounds on typing speeds using a stylus and soft keyboard, *Behaviour & Information Technology 14* (1995), 370-379.
21. Yamada, H. A historical study of typewriters and typing methods: From the position of planning Japanese parallels, *Journal of Information Processing 2* (1980), 175-202.
22. Zhai, S., Hunter, M., and Smith, B. A. The Metropolis keyboard: An exploration of quantitative techniques for graphical keyboard design, *Proceedings of the ACM Symposium on User Interface Software and Technology - UIST 2000*. New York: ACM, 2000, 119-128.
23. Zhai, S., Hunter, M., and Smith, B. A. Performance optimization of virtual keyboards, *Human-Computer Interaction 17* (2002), 229-269.
24. Zhai, S., Sue, A., and Accot, J. Movement mode, hits distribution and learning in virtual keyboarding, *ACM Conference on Human Factors in Computing Systems -- CHI 2002*. New York: ACM, 2002, 17-24.