INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION, *16*(2), 163–184

Data Entry for Mobile Devices Using Soft Keyboards: Understanding the Effects of Keyboard Size and User Tasks

Andrew Sears
Ying Zha
Interactive Systems Research Center, Information Systems Department
UMBC

As mobile, handheld computing devices become more common and are used for an ever-increasing variety of tasks, new mechanisms for data entry must be investigated. Personal digital assistants often provide a small stylus-activated soft keyboard, as do some mobile phones that include touch screens. However, there is little data regarding the importance of keyboard size or the users' tasks, the effectiveness of these keyboards, or user reactions to these keyboards. In this article, an experiment designed to investigate these issues in the context of a palm-style QWERTY keyboard is described. In this study, 30 novices completed 6 realistic tasks using either a small, medium, or large soft keyboard. The results not only confirm that keyboard size does not affect data entry rates but that making the keyboard smaller does not increase error rates or negatively impact preference ratings. However, tasks that required users to switch between the alphabetic keyboard and the numeric keyboard do result in significantly slower data entry rates. A model that accurately predicts the time required to enter predefined text is presented, and directions for future research are discussed.

1. INTRODUCTION

Mobile, handheld computing devices are becoming increasingly common. Personal digital assistants (PDAs), mobile phones, and pagers support an ever expanding array of activities. While users continue to record names, addresses, appointments, and reminders, they are also using these devices to access both e-mail and the World Wide Web.

As the range of applications continues to expand, data entry is becoming more important. The most significant increase in data entry activities may be associated

---

This research was funded by Motorola. We gratefully acknowledge their generous support that made this research possible.

Requests for reprints should be sent to Andrew Sears, Interactive Systems Research Center, Information Systems Department, UMBC, 1000 Hilltop Circle, Baltimore, MD 21250. E-mail: asears@umbc.edu

with generating replies to text pages and e-mail messages. To support these activities, mobile devices employ a variety of data entry techniques including stylus-activated soft keyboards (e.g., a small QWERTY keyboard presented on a touch-sensitive screen), small physical keyboards, stylus-based gesture recognition (e.g., Jot and Graffiti; see Sears & Arora, in press), and telephone keypad-based techniques (e.g., T9 and multitap). Although some studies have been reported that explored these techniques (see MacKenzie & Soukoreff, 2002, for an overview), many questions have yet to be answered. In an earlier study, Sears and Arora (2002) investigated the efficacy of Jot and Graffiti. This study builds on the study by Sears and Arora, using the same experimental methodologies, to investigate the efficacy of stylus-activated soft keyboards.

Although stylus-activated soft keyboards are standard on most PDAs and are available on some mobile phones, limited data exists regarding the efficacy of this common data entry technique. Further, many existing studies have employed simple tasks, desktop tablets, or unrealistic strategies for addressing errors. Therefore, the results of these studies may not accurately reflect the performance or satisfaction that should be expected when stylus-activated soft keyboards are used to complete realistic tasks using handheld devices such as a PDA or mobile phone. Given these concerns, in this study, we investigated the efficacy of a QWERTY-style, stylus-activated soft keyboard for completing a variety of realistic tasks when using a PDA.

Whereas PDAs often employ soft keyboards that are as wide as 5.5 cm, mobile phones often have displays as narrow as 3.0 cm. Fitts's (1954) Law, which has been applied to both soft keyboards and small devices, would predict that reducing the size of the keyboard would not affect data entry rates. The logic is that as long as the distance between the keys (D) is reduced at the same rate as the size of the keys (W), the ratio of D:W remains constant, the index of difficulty remains constant, and data entry rates will be unchanged. However, it is unclear whether Fitts's Law is sufficient for evaluating the efficacy of small stylus-activated keyboards such as those used in this study. There are three reasons why predictions based on Fitts's Law may not be sufficient.

First, Fitts's Law has not been validated in the context of small soft keyboards. Some studies have predicted data entry rates but did not provide formal validation of these predictions (e.g., Silfverberg, MacKenzie, & Korhonen, 2000), most used keyboards that were larger than those used in this study (e.g., Mackenzie & Zhang, 2001; Soukoreff & MacKenzie, 1995), and others did not indicate the size of the keyboards being used (e.g., MacKenzie, Nonnecke, Ridersma, McQueen, & Meltz, 1994; MacKenzie, Zhang, & Soukoreff, 1999). To date, no studies have been reported using soft keyboards as small as those employed in this study, data entry rates were predicted using Fitts's Law, and these predictions were empirically validated.

Second, several studies have suggested that Fitts's Law may not be the most appropriate method of assessing the efficacy of small keyboards such as those used in this study. For example, Sears, Revis, Swatski, Crittenden, and Shneiderman (1993) found that reducing the size of soft keyboards resulted in slower data entry rates, more corrected errors, and lower satisfaction ratings (a more detailed discussion of this article is provided later). Although participants entered data by touching a soft keyboard displayed on a touch screen using their finger, and the keyboards were larger than those used in this study, these results suggest that Fitts's Law may not be sufficient for evaluating the efficacy of small soft keyboards. A recent example is by James and Reischel (2001), who provided empirical data on data entry rates using mobile phones that are inconsistent with the Fitts's Law-based predictions reported by Silfverberg et al. (2000). Although the predictions and empirical data were based on different mobile phones, the small differences in the physical devices cannot explain the much larger differences between the predicted and observed data entry rates.

Third, Fitts's Law focuses exclusively on the time required to complete physical movements. Fitts's Law does not consider mental preparation time or visual scanning time, assumes error-free performance, and provides no insights with regard to user preferences. It is possible that physical movements may not take longer, but the perceptual demands would increase the time required as the keys became smaller. It is also possible that smaller keyboards may be just as fast but that error rates will increase or that users would simply prefer larger keyboards.

Even if Fitts's Law is shown to accurately predict performance with small soft keyboards, questions would remain regarding error rates and user perceptions. Therefore, this study was designed to provide a comprehensive understanding of the efficacy of small, stylus-activated, soft keyboards by investigating data entry rates, error rates, and user preferences. As part of this process, we reassessed the efficacy of Fitts's Law for evaluating soft keyboards.

These issues were investigated with an emphasis on novice users. Novice performance and preferences are considered important for two reasons. First, many individuals make decisions regarding the adoption of these technologies after limited interactions. Therefore, initial performance and preferences may affect whether or not a technology is adopted for future use. Second, many individuals use these technologies only occasionally. Therefore, their performance is likely to mimic that of novices rather than well-trained experts.

The results of this study should prove useful to both product designers and researchers. For product designers, the results provide insights that will allow for more informed decisions regarding which technology to adopt. The results also identify potential problems and therefore highlight opportunities for additional research. Of course, keyboard size is only one of many issues that must be considered. For example, Sears, Jacko, Chu, and Moro (2001) highlighted the importance of the users' familiarity with the organization of the letters within the keyboard and the number of letters located on each key.

2. RELATED RESEARCH

Numerous studies have been conducted with the goal of better understanding the efficacy of stylus-activated, QWERTY-style keyboards. Most, if not all, of these studies were motivated by the fundamental problem of entering data on small, handheld, mobile devices.

Perhaps the most recent study, by MacKenzie and Zhang (2001), explicitly addressed some of the questions of interest in this research. In this study, we employed two different key sizes (10.0 × 10.0 mm and 6.4 × 6.4 mm) and two different keyboard layouts (QWERTY and a random layout in which letters were rearranged after each keystroke). Participants entered 10 phrases of approximately 25 characters using each of the four possible combinations of keyboard size and layout. When using the QWERTY layout, participants were able to enter text at approximately 20 to 21 words per minute (wpm) as compared to approximately 5.5 wpm when using the random keyboard layout. Clearly, familiarity with the keyboard layout allows for more effective data entry. As would be predicted by Fitts's Law, keyboard size did not have a significant effect on data entry rates. Interestingly, the random keyboards resulted in significantly lower error rates, whereas the smaller keyboards resulted in significantly higher error rates. As with a number of the studies described following, the keyboard was presented on a Wacom tablet, only uppercase characters were required to complete the tasks, and participants were instructed not to correct errors.

Sears et al. (1993) investigated the relation between the size of a touch screen-based soft keyboard and data entry rates. Participants used four soft keyboards that varied in size to enter several simple phrases. The keyboards ranged from 6.8 to 24.6 cm wide (measuring between the outer edges of the "Q" and "P" keys). Novices could enter text at 10 to 20 wpm depending on the size of the keyboard, whereas experienced users could achieve data entry rates ranging from 21 to 32 wpm. Perhaps the most important results are those that highlight a relation between keyboard size and the various dependent variables that were measured. Users made and corrected fewer errors when entering text on larger keyboards, data entry rates were higher on larger keyboards, and participants preferred larger keyboards.

MacKenzie and Zhang (1999) conducted an experiment with the goal of identifying the upper limit for data entry rates for a QWERTY-style soft keyboard and a new optimized keyboard layout. Participants completed 20 sessions during which they used each keyboard layout for 22 min. As a result, this study provides data about both immediate use and practiced performance. During the sessions, participants used a Wacom tablet to enter predefined phrases composed of uppercase letters. No details were provided regarding the size of the keyboards. Participants did not correct errors, but audible feedback was provided whenever an incorrect letter was entered and participants were instructed to slow down if their error rate exceeded 10% for any given trial. Performance on the QWERTY keyboard ranged from 28 wpm for the first trial to approximately 40 wpm by end of the study.

MacKenzie et al. (1999) described a study comparing six alternative keyboard layouts. Participants tapped on a paper image of the keyboard to enter one predefined phrase consisting of lowercase letters. Participants were instructed to complete the task as quickly as possible while avoiding errors. Error rates were not recorded and no details were provided regarding the size of the keyboards. When using the QWERTY layout, participants were able to enter more than 20 wpm.

MacKenzie, Nonnecke, McQueen, Riddersma, and Meltz (1994) investigated three alternatives for entering text on pen-based computers—including a stylus-activated, QWERTY-style soft keyboard. Participants used a Wacom tablet to enter 22 character phrases composed only of lowercase letters. Participants were instructed to aim for both speed and accuracy but were also instructed to ignore mistakes. When using the QWERTY keyboard, participants were able to enter over 22 wpm.

Lewis, LaLomia, and Kennedy (1999) had participants enter sentences using six paper mockups of several alternative keyboard layouts. Participants were instructed to enter sentences as quickly and accurately as possible. When errors occurred, participants were instructed to enter the correct letter (without deleting the incorrect letter) and to continue. Data entry rates for the QWERTY layout reached approximately 24 wpm.

Lewis (1999) also compared three alternatives for data entry on handheld devices including a stylus-activated, QWERTY-style soft keyboard. Participants used a Simon PDA to enter both addresses and sentences. Participants were required to produce 100% accurate text by correcting errors, but the procedure for verifying the accuracy of the results before allowing a participant to continue was not specified. Data entry rates ranged from approximately 11 wpm for addresses to 17 wpm for sentences.

These studies were motivated by the fundamental problem of entering data on small, handheld, mobile computing devices. Each study included a QWERTY-style keyboard as one of the alternatives explored. Data entry rates ranged from 11 to 28 wpm for initial performance and as high as 40 wpm after extensive practice, but the methodologies and tasks employed make it difficult to generalize many of these results to situations in which individuals are completing realistic tasks on a PDA. Some studies used paper mockups of the keyboards, others used desktop tablets, and the one study that did use a PDA utilized a soft keyboard almost twice as large as those found on most PDAs used today. Several studies had users either ignore errors or type the correct letter without deleting the incorrect letter. These strategies may provide insights into the data entry rates that may be possible when users do not make mistakes but do not allow us to determine the data entry rates that are possible under realistic conditions in which users do correct errors. The most representative results are those reported by Lewis (1999), but the keyboard used by Lewis was over 10.0 cm wide—leaving the fundamental questions raised earlier unanswered.

3. RESEARCH OBJECTIVES

Soft keyboards are readily available on PDAs, yet there is little knowledge regarding their effectiveness for novice users. Previous studies have provided some insights, but the experimental materials and procedures, as well as the tasks utilized in these earlier efforts, may limit the application of the reported results when dealing with novices completing realistic tasks using a PDA. Therefore, the first objective of this study was to gain additional insight into the effectiveness of a soft keyboard under these conditions. To do this, data entry rates, the number of uncorrected errors, and the satisfaction of study participants are reported.

Soft keyboards have not been widely adopted as an alternative for data entry on mobile phones or other devices with smaller displays than those on typical PDAs.

Although Fitts's Law may predict that changing the keyboard size will have no effect on performance, it only addresses physical movements. Fitts's Law does not address the time spent visually scanning a display to locate the desired key, transitions between alternative keyboards, or delays that occur as users first begin their task. Further, Fitts's Law itself has never been validated in the context of data entry using a stylus-activated soft keyboard on a PDA. Finally, even if performance is not affected by keyboard size, reducing the size of the keyboard may result in more negative subjective ratings by users. The second objective of this study was to determine whether or not keyboard size affects performance or preferences in the context of data entry activities using stylus-activated soft keyboards that would fit on a PDA or mobile phone. As a result, the keyboards used were substantially smaller than those used in earlier studies.

Although alternative layouts continue to be explored, the QWERTY layout remains the most prevalent design for both physical and soft keyboards. Even after the QWERTY layout is chosen as the basic design for a soft keyboard, additional decisions still remain. For example, both Palm® and Casio devices employ soft keyboards based on the QWERTY layout. Casio devices provide a keyboard that closely mimics traditional physical keyboards, providing quick access to all of the standard characters. Palm devices use a simplified QWERTY layout as the primary keyboard, with alternative keyboards providing access to numbers, additional symbols, and international characters. As a result, fewer keys are present on each Palm keyboard, but each key is larger. Given the limited screen space available on devices such as mobile phones, the trade-off between having more small keys or fewer keys that are larger becomes critical and should be the focus of multiple studies. In this study, the effectiveness of a Palm-style keyboard was evaluated (see Figure 1). Although only one keystroke is required to switch key-

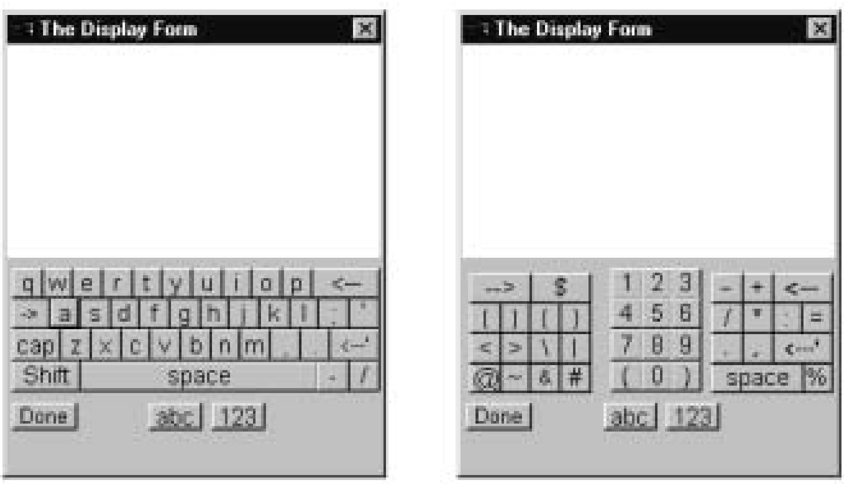

FIGURE 1 The large two-screen QWERTY style soft keyboard used in this study. The "abc" and "123" keys change the display to the appropriate keyboard.

boards, users must determine that it is necessary to switch keyboards and then reorient themselves once the transition is complete. The time involved in switching between keyboards will be important when such transitions become frequent. Given the importance of keyboard transitions, we believe that existing models for data entry rates are inadequate. Therefore, the final objective was to propose a model for data entry rates for keyboards like those utilized in this study. To do this, the approach used by Card, Moran, and Newell (1980) in the keystroke level model (KLM) was utilized.

4. METHOD

4.1. Participants

Thirty students enrolled at UMBC volunteered to participate in the study and received a payment of $10.00 as compensation for their time. To better represent the potential users of these techniques, participation was restricted to students who were not in an information technology or engineering-based major (e.g., Information Systems, Computer Science, or Computer Engineering). Informed consent was obtained prior to participation. A between-group design was used for this study, with participants randomly assigned to use one of three different sized keyboards when completing their tasks. Ten participants used each keyboard size.

Demographic information, including age, gender, and computer experience, was gathered at the conclusion of the study. Seventeen participants were women. The average age of the participants was 25.6 ($SD = 9.5$). All participants were regular computer users. Ten participants used mobile phones an average of 1.8 times per day. Seven participants (including 3 of the mobile phone users) used pagers an average of 3.7 times per day. Sixteen participants did not use mobile phones or pagers. None of the participants used PDAs.

4.2. Apparatus

A Casio Cassiopia E100 (Casio, Inc., Dover, NJ) was used in this study. The Casio Cassiopia E100 measures approximately 8.4 × 13.0 × 2.9 cm and weighs approximately 255 g. The size of the display/input region is 6.1 × 7.9 cm. Data entered by the participants was displayed using the top of this region, with the bottom of the display being reserved for the soft keyboard. The primary (alphabetic) keyboard was either 3.2 cm, 4.3 cm, or 5.4 cm wide. Throughout the remainder of this article, these keyboards are referred to as small, medium, and large, respectively. Initially, the large keyboard was created. Next, the small and medium keyboards were created by reducing the size of each key. Alphabetic keys were square, measuring approximately 2.6 mm, 3.5 mm, and 4.4 mm per side for the small, medium, and large keyboards, respectively. The tops of the three keyboards were located at the same position on the screen and each keyboard was centered horizontally. Although earlier studies have provided mixed results when comparing the lift-off, first-contact, and land-on strategies, the lift-off strategy was selected for this study because it is more consistent with the way soft keyboards are currently implemented on most mobile devices. When using the lift-off selection strategy, the key that is highlighted when the stylus is lifted from the screen is selected (Potter, Weldon, & Shneiderman, 1988).

The small keyboard was the smallest keyboard that could be displayed given the screen resolution and available fonts. This keyboard could fit on many, but not all, mobile phones. The large keyboard represented a typical size for a soft keyboard on a PDA—using virtually the entire width of the display.

To maximize the size of each key, the two-screen, QWERTY-style keyboard used in Palm devices was utilized (see Figure 1). This design provides access to all letters and some symbols from the primary keyboard (on the left) but requires users to access a secondary keyboard (on the right) to type numbers and some other symbols. As a result, the transition from the main keyboard to the secondary keyboard is likely to prove important when tasks involve a mixture of letters, numbers, and symbols. All participants' interactions were recorded by the software being utilized. The keyboards were implemented using Visual Basic® and the Windows® CE Toolkit for Visual Basic Version 6.0.

4.3. Tasks

The keyboards employed in this study require users to switch between the primary and secondary keyboards when tasks involve numbers and some symbols. Such tasks would be expected to result in slower data entry rates. Similarly, uppercase letters are expected to require more time because an extra keystroke is required (i.e., the Cap lock key or Shift key). Therefore, six tasks that differed in length and content were utilized, and task design was guided by a fundamental goal of engaging the study participants in tasks that are representative of those they would encounter when using an Internet enabled mobile device.

Task 1 involved entering a name and address as may be done when completing an entry in an application designed to keep track of contact information (e.g., an address book). This task was selected due to the various numbers and symbols that would be required when recording detailed contact information. Tasks 2 and 3 involved entering URLs. Relatively simple URLs were utilized because more complex URLs are likely to be bookmarked or obtained by entering a simple URL and subsequently navigating to the desired location. URLs are unique in that they are composed of a sequence of letters, numbers, and symbols without intervening spaces. Tasks 4 through 6 involved entering varying amounts of basic alphanumeric data, but unlike Task 1, the use of symbols and formatting was minimal. Task 4 involved entering two words as may be done when recording the topic of an upcoming appointment. Task 5 involved entering a short sentence, which could correspond to a longer description of a meeting or a brief reply to an e-mail or page. Task 6 involved entering a single paragraph as may be done when responding to an e-mail message.

As illustrated in Table 1, a total of 50 unique numbers, symbols, and letters were entered during the six tasks. This included the digits 0 through 9, eight different symbols/punctuation marks, nine uppercase letters, and 23 lowercase letters. The exact text for each task is included in the Appendix.

Comparing how frequently each letter is used in written English (Pratt, 1939) and how frequently each letter is used during the six tasks results in a Pearson correlation coefficient of $r = .92$ ($p < .001$). Although the frequencies are not identical, letters that occur frequently in written English occur frequently in these tasks and letters that occur infrequently in written English occur infrequently in these tasks.

4.4. Dependent Variables

Dependent variables were defined to allow performance, satisfaction, and the underlying process to be assessed. For performance, the focus was on data entry rates—measured in corrected wpm and uncorrected errors. To account for variable word lengths among tasks, the five-character word counts from the last column of Table 1 were used throughout the data analysis. The time to complete each task was converted into a data entry rate (corrected wpm). As would be expected during realistic usage, participants corrected most but not all errors. Therefore, uncorrected errors were identified by comparing the desired result (see Appendix) to the actual text produced by each participant. The number of errors was divided by the number of words in the task to determine the uncorrected error rates. The number of characters participants corrected and the number of correction sequences were also extracted from the interaction logs generated by the keyboard software.

For satisfaction, a questionnaire was administered that investigated feelings regarding how easy it was to enter text, how quickly text could be entered, and the acceptability of the accuracy of the technique. The questionnaire also investigated whether the individual was comfortable using the device, felt physically tired when using the device, or would be interested in using the device in the future. All

Table 1: Characteristics of Tasks One Through Six

Task	*Total Characters*[a]	*Unique Characters*[b]	*Actual Words*[c]	*Words for WPM*[d]
1	86	40	13	17.2
2	18	14	1	3.6
3	26	15	1	5.2
4	18	12	2	3.6
5	44	16	8	8.8
6	223	28	41	44.6
Total	410	50	75	83.0

Note. WPM = words per minute.

[a]Character count includes spaces. Required carriage returns are included for task one. [b]Each unique uppercase and lowercase letter, number and symbol/punctuation mark was counted once. [c]Number of actual words (computed using spaces to define word boundaries). [d]Words for wpm count based upon a frequently used standard for written English (Gentner, 1983) of 5 characters/word (4 letters plus a space). Average word lengths differ in other languages.

responses were provided using a scale ranging from 1 (*strongly agree*) to 5 (*strongly disagree*). The composite reliability of this questionnaire, which has been used in previous studies, assessed through use of Cronbach's α, was 0.91. The robustness of this result confirms the internal consistency of the survey. In addition, a variety of measures were used to allow a detailed investigation of the process users employed when completing their tasks.

4.5. Procedure

Keyboard size was treated as a between-group variable with each participant interacting with either the small, medium, or large keyboard. A between-group design was used for two reasons. First, this allowed the same tasks to be completed with all three keyboards. Even a counterbalanced within-subjects design would not have accomplished this goal due to the learning that would have occurred as participants were exposed to the text they were required to enter. Second, a goal was to understand how the participants would react to using each keyboard when they were not aware that keyboard size was being varied. A within-subjects design would result in participants using all three keyboards and this could inappropriately affect their preference ratings. Task was treated as a within-subjects variable with each participant completing all six tasks in a unique random order.

After reading and signing a consent form, participants were given the device they would be using (turned off). At this time, they were asked to write down the types of activities they would expect the device to support. Next, they provided similar information by selecting activities from a predefined list. This information was gathered as part of a long-term project exploring the relation between the physical characteristics of handheld devices and perceived uses. Because only one device was used in this study, these data were not expected to provide any insights at this time.

Throughout the study, participants were free to hold the device in their hand or rest it on a table. Participants were provided with a brief orientation to the device. This orientation included a demonstration of how to use the keyboard including how to switch between the primary and secondary keyboards. Next, participants were allowed 10 min to practice entering data. To structure their practice, they were given a collection of example tasks that were similar in length and content to the actual tasks they would be assigned. Participants were free to practice using these sample tasks or any other text.

Participants were provided with one task at a time to ensure that the tasks were completed in the appropriate order. For each task, the participant was given a sheet of paper with the required text. They were allowed to review the text and begin the task when they were ready. Participants were instructed to complete the task as they would under realistic usage conditions, balancing speed and accuracy. As a result, participants corrected most but not all errors. The experimenter did not provide help during the experimental sessions. Participants were allowed to take a break before beginning each task. These breaks averaged approximately 30 sec. All participant interactions were automatically recorded with timestamps by the keyboard software.

After completing all six tasks, participants completed two additional questionnaires. The first investigated participants' perceptions of the technique (i.e., the soft keyboard) they used. The second gathered basic demographic information.

5. RESULTS

Preliminary data analysis explored both productivity (i.e., data entry rates and uncorrected errors) and user perceptions (i.e., data from subjective satisfaction questionnaires). More detailed analysis explored the process by which participants entered data (i.e., the number of characters they corrected, the number of correction sequences, the delay before entering the first character, and the delay involved in switching between the primary and secondary keyboards).

5.1. Data Entry Rate (Corrected WPM)

Means and standard deviations for data entry rates are reported in Table 2. A two-way mixed effects analysis of variance (ANOVA) with keyboard size as a between-group variable and task as a within-subjects variable was used to assess the effect of keyboard size. There was no significant main effect for keyboard size, $F(2, 27) = 0.59$, $p > .55$, but there was a significant effect for task, $F(5, 135) = 50.22$, $p < .001$. Because keyboard size did not affect data entry rates, average data entry rates across all three keyboard sizes are illustrated in Figure 2.

To provide a more comprehensive understanding of these results, the counternull value was computed (Rosenthal & Rubin, 1994). Although the counternull value is not widely reported in the human–computer interaction literature, it is particularly valuable when the null hypothesis is not rejected, as it provides insights into the magnitude of the effect that may exist. More specifically, by comparing counternull values for different conditions, bounds on the size of the effect keyboard size had on data entry rates could be placed instead of simply assuming (incorrectly) that there was no effect at all. Because no standardized notation exists for representing counternull values, these values can be viewed as alternatives to the empirically measured means, and when reporting values for small, medium, and large keyboards, the following notation was used: CNM_S, CNM_M, and

Table 2: Mean Data Entry Rates (in Words Per Minute) and Standard Deviations for Each of Six Tasks

	Task													
	1		*2*		*3*		*4*		*5*		*6*		*Overall*	
Keyboard Size	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Small	7.49	0.99	7.60	1.99	9.33	1.82	9.66	2.98	10.39	3.20	12.26	2.81	9.46	2.85
Medium	7.25	1.40	6.77	1.01	7.56	1.18	7.67	2.73	10.20	2.84	12.56	2.71	8.67	2.91
Large	6.98	0.85	7.31	1.59	9.33	2.07	8.53	2.31	10.49	2.74	12.62	1.58	9.21	2.70

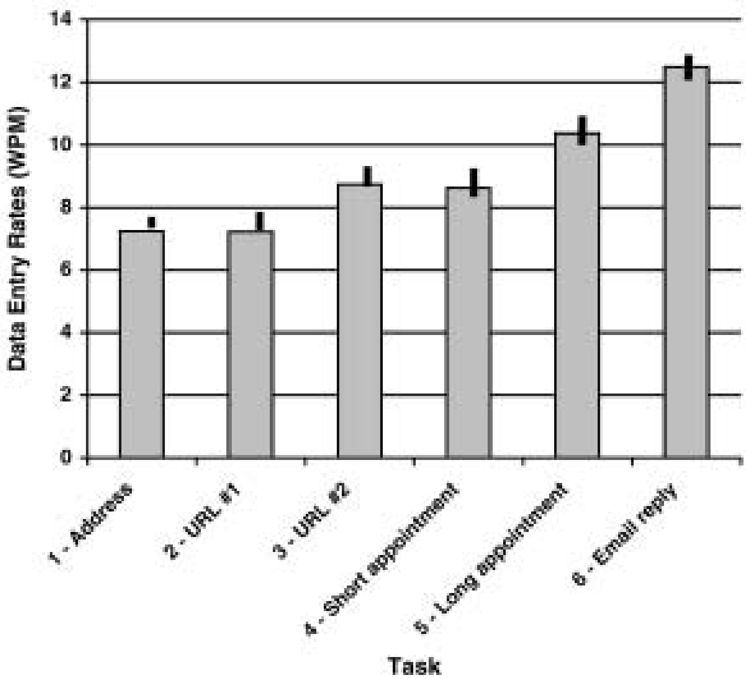

FIGURE 2 Average data entry rates for each of six tasks (bars indicate 95% confidence interval). WPM = words per minute.

CNM_L represent the counternull means for the small, medium, and large keyboards, respectively ($CNM_S = 9.80$, $CNM_M = 8.22$, and $CNM_L = 9.31$). Instead of incorrectly assuming that all three keyboards resulted in the same data entry rates, the counternull values allow the following statements to be made:

- The small keyboard resulted in performance that was somewhere between 0% and 19% faster than the medium keyboard.
- The small keyboard resulted in performance that was somewhere between 0% and 5% faster than the large keyboard.
- The medium keyboard resulted in performance that was somewhere between 0% and 12% slower than the large keyboard.

The lack of a significant difference suggests that changing the keyboard size did not result in any significant changes in data entry rates. The counternull analysis further suggests that changing the size of the keyboard had an effect that varied between 0% and 19%. However, in this specific situation, the detailed results of the counternull analysis are particularly interesting. Note that moving from the large to medium keyboard results in no change or slower performance, whereas moving from the large to small keyboard results in no change or faster performance. Because there is no logical reason why a small reduction in keyboard size should hinder performance, and an even larger reduction should prove beneficial, these results suggest that keyboard size is unlikely to have a practical impact on data entry rates under the conditions employed in this study.

Planned comparisons for the effect of task on data entry rates were performed using contrast matrices. The significant results can be summarized as follows (T = task): T1 = T2 < T3 = T4 < T5 < T6 ($p < .02$ for all comparisons). In other words, Tasks 1 and 2 resulted in equivalent data entry rates, as did Tasks 3 and 4, but all other comparisons were significant. These differences are explored in more de tail following through the analysis of the process users employed when interacting with the soft keyboards.

5.2. Uncorrected Error Rates

Means and standard deviations for uncorrected error rates are reported in Table 3. A two-way mixed effects ANOVA with keyboard size as a between-group variable and task as a within-subjects variable was used to assess the effect of keyboard size. There was no significant main effect for keyboard size or task, $F(2, 27) = 0.99$, $p > .38$ and $F(5, 135) = 0.55$, $p > .74$, respectively. Although keyboard size did not significantly effect uncorrected error rates, an analysis of counternull values indicated that reducing the size of the keyboard (from large to small) reduced errors by somewhere between 0% and 15%.

5.3. Preferences

After completing all six tasks, participants completed a questionnaire. The first four questions inquired about how easy it was to enter addresses (Task 1), URLs (Tasks 2 and 3), appointments (Tasks 4 and 5), and e-mail replies (Task 6) tasks. Questions 5 through 8 investigated perceived speed for completing these tasks. Questions 9 through 12 investigated the participants' perceptions of the number of errors made during the tasks. Questions 13 through 15 investigated perceived comfort, feelings of being physically tired, and interest in using the device in the future, respectively. All responses resulted in a score between 1 and 5, with 1 being the most positive response for each question. Reliability was assessed using

Table 3: Means and Standard Deviations for Uncorrected Errors Rates for Each of Six Tasks

	Task													
	1		*2*		*3*		*4*		*5*		*6*		*Overall*	
Keyboard Size	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Small	0.052	0.108	0.056	0.117	0.039	0.122	0.028	0.088	0.011	0.036	0.038	0.083	0.034	0.090
Medium	0.000	0.000	0.028	0.088	0.000	0.000	0.000	0.000	0.000	0.000	0.036	0.043	0.011	0.041
Large	0.017	0.028	0.028	0.099	0.019	0.061	0.056	0.176	0.034	0.077	0.047	0.042	0.037	0.094

Cronbach's alpha. The resulting value of $\alpha = 0.83$ indicates that the questionnaire is reliable and substantiates the robustness of these results.

A two-way mixed effects ANOVA with keyboard size as a between-group variable and question as a within-subjects variable was used to assess the effect of keyboard size. Although there was no significant main effect for keyboard size, $F(2, 27) = 0.49$, $p > .61$, there was a significant effect due to the question asked, $F(14, 378) = 4.45$, $p < .001$. Interestingly, planned comparisons of responses for individual questions yielded only one significant difference. Keyboard size had a significant effect on participant responses to the question regarding the ease of entering URLs, $F(2, 27) = 8.79$, $p < .002$, but no consistent pattern could be identified between keyboard size and participant responses (i.e., mean responses were 2.8, 1.5, and 2.1 for the small, medium, and large keyboards, respectively).

5.4. Process

To better understand the process participants employed while completing these tasks, two additional sets of analyses were performed. The first focused on how users corrected errors (i.e., corrected characters and correction sequences). The second focused on providing the information necessary to effectively model task completion times. More specifically, the time required to enter the first character when beginning a new task was identified, as well as the time involved in switching between the two keyboards.

Corrected characters and correction sequences. Analyses identical to those used for uncorrected error rates were performed on the number of characters corrected while completing the tasks as well as the number of correction sequences. A *correction sequence* was defined as a collection of consecutive backspaces (e.g., five consecutive backspaces were counted as one correction sequence).

There was no significant main effect for keyboard size on either corrected characters or correction sequences, $F(2, 27) = 0.52$, $p < .59$, and $F(2, 27) = 0.05$, $p > .95$, respectively. Further, there was no significant effect for task on either corrected characters or correction sequences, $F(2, 27) = 1.77$, $p > .12$, and $F(2, 27) = 1.97$, $p > .08$, respectively.

Start-up time and keyboard transitions. Not all keystrokes are the same. To demonstrate these differences, keystrokes were divided into four groups. The first group consists of the first character of each task. When users began a new task, they had to orient themselves to the screen. Therefore, the first character was expected to take longer than subsequent characters. This start-up time will be particularly important for novices who have less experience with an interface and when tasks involve entering small quantities of text.

When the user needed a character that was not available on the current keyboard, they had to select either the "abc" or "123" key. Before initiating this transition, the user had to confirm that the desired character was not available on the current keyboard. After selecting the "abc" or "123" key, the user had to reorient themselves to the new keyboard. Therefore, the second group of keystrokes included every selection of the "abc" or "123" keys (i.e., after confirming that they had to change keyboards). The third group of keystrokes included the keys selected immediately following the "abc" or "123" key (i.e., after reorienting themselves). The fourth group included all key presses not included in any of the other three groups, including the shift key when uppercase letters were required.

To confirm that entering the first character (Group 1) required longer than the characters that followed (Group 4), the average time for characters in Groups 1 and 4 were compared. As expected, a paired sample t test identified a significant difference between the start-up (M = 4.11 sec) and key-press times (M = 1.07 sec), $t(29) = 8.1$, $p < .001$.

Each transition from one keyboard to the other involves a decision (Group 2 keystroke) and a recovery (Group 3 keystroke). Whereas the decision is associated with a key press that is required only because there are multiple keyboards, the recovery is associated with a key press that would be required even if there were only a single keyboard. Therefore, all of the decision time is associated with the transition, but only part of the recovery time is allocated to the transition. More specifically, the additional time that is introduced as a result of requiring a transition between keyboards can be defined as follows:

$$\text{transition time} = \text{decision time} + (\text{recovery time} - \text{key-press time})$$

The average transition time was computed for each participant. As expected, a paired sample t test identified a significant difference between the transition (M = 3.30 sec) and key-press times (M = 1.07 sec), $t(29) = 14.8$, $p < .001$. The mean start-up, key-press, and transition times are used following to model overall data entry rates.

6. MODELING DATA ENTRY RATES

Earlier attempts to model data entry rates for soft keyboards often employed Fitts's Law to predict the time required to select a predefined sequence of characters (e.g., Soukoreff & MacKenzie, 1995). This approach assumes that the distance traveled is the only difference between keystrokes. Some models also integrated a simple visual search component based on the Hick–Hyman (Hick, 1952; Hyman, 1953) model of choice reaction times (e.g., Soukoreff & MacKenzie, 1995). However, results of a recent study by Sears et al. (2001) suggest that Hick–Hyman is inappropriate for this task. Using Hick–Hyman implies that only the number of keys is important when determining which key to press. In contrast, Sears et al. (2001) provided evidence that both the keyboard layout (e.g., QWERTY, Dvorak) and the number of letters represented by each key (e.g., three per key on a telephone keypad) must be considered.

Existing models based on Fitts's Law do not address the time involved in moving between alternative keyboards or the additional time required to enter the first character when starting a new task. Further, there appear to be fundamental problems with both Fitts's (1954) Law and the Hick–Hyman (Hick, 1952; Hyman, 1953)model of choice reaction time in the context of small stylus-activated soft keyboards. The use of the Hick–Hyman model for visual search has been shown to be inappropriate by both Sears et al. (2001) and MacKenzie and Zhang (2001). The keystroke-level analysis presented previously provides the first empirical evidence that Fitts's Law is not appropriate for modeling user interactions with soft keyboards like those used in this study. The failure of Fitts's Law to accurately model the start-up time is most important when a limited number of characters are entered (e.g., Tasks 2 and 4). The failure to accurately model keyboard transitions becomes important in situations like Tasks 1, 2, and 3 in which characters are required that are not available on the primary keyboard. Therefore, we propose a KLM-style model that shifts the focus from predicting the time required to move between specific keys to predicting the total time necessary to complete tasks when multiple characters are entered. The model builds on the following definitions:

T = total time to complete the task.
t_1 = time for the first key press when beginning a new task.
t_d = time to make a decision that a transition is required.
t_r = time to recover from a transition and complete the subsequent key press.
t_k = time for each additional keystroke (not addressed by t_1, t_d, or t_r).
c = number of characters required by the task.
c_s = number of shifted characters (e.g., uppercase letters or alternative symbols).
c_t = number of transitions between keyboards required by the task.

Given these definitions, the total task completion time is defined as

$$T = t_1 + [(t_d + t_r)c_t] + [t_k(c + c_s - c_t - 1)]$$

Figure 3 presents the predicted data entry rates that result from using the following experimentally derived values: $t_1 = 4.11$, $t_d = 1.83$, $t_r = 2.47$, and $t_k = 1.07$. The c, c_s, and c_t were computed based on the text that was entered for each task (see Appendix). A strong correlation existed between the predicted times and average of the results across all of the participants (Pearson correlation coefficient $r = .99$). As would be expected, using the predictions (and actual results) for each individual participant resulted in a lower correlation, but the relation was still strong (Pearson correlation coefficient $r = .94$). Therefore, the proposed model can be used to make useful predictions regarding the total time required to complete a task. The key difference between the proposed model and existing models is the explicit acknowledgement that the initial keystroke, uppercase letters, and transitions between keyboards all take longer than the normal keystrokes required to enter a single lowercase letter. Unlike models that assume that all keystrokes are the same, the proposed model predicts different data entry rates based on the number of uppercase letters and keyboard transitions required.

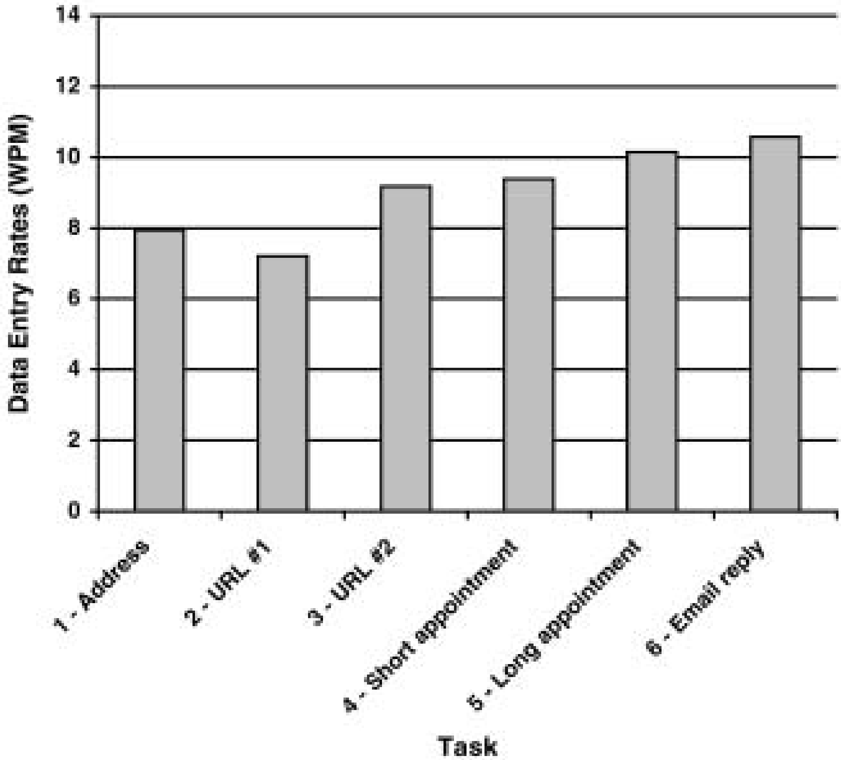

FIGURE 3 Predicted data entry rates for each of six tasks. WPM = words per minute.

This model can be used to predict the effects of switching from a Palm-style, two-keyboard design to a design in which all keys are available on a single keyboard. Clearly, the primary benefit of this change would be the elimination of the transitions between keyboards. All other aspects of the keyboard would remain the same. Therefore, the time required to complete the tasks can be modeled by simply setting $c_t = 0$ for all tasks. Because Tasks 4 and 5 did not involve any keyboard transitions, data entry rates remained unchanged. Task 6 involved two transitions, but given the length of this task, eliminating the keyboard transitions had no practical impact on data entry rates. The benefits of a single keyboard design became more apparent for Tasks 1 through 3 in which users had to switch between the keyboards several times. The predicted data entry rates were approximately 25%, 28%, and 10% higher for the single keyboard design as compared to the Palm-style, two-keyboard design. However, much of this gain could be realized by an intermediate design in which some additional keys are made available on the primary keyboard with other, less frequently used keys still being placed on the secondary keyboard. For example, simply making the numbers available on the primary keyboard could provide substantial benefits. The model described previously could be used for preliminary evaluations of alternative keyboard designs before resources are committed to implementation and formal evaluations.

7. DISCUSSION

Overall, keyboard size had no significant effect on performance (i.e., data entry rates or uncorrected errors). Similarly, responses to only one of the 15 questions on the satisfaction questionnaire were affected by keyboard size, and even those results did not show a consistent pattern. The importance of these results are clarified by revisiting the original motivation for this research.

First, it is becoming increasingly common for mobile computing devices to be used for data entry activities. As a result, a more comprehensive understanding of the benefits and limitations of existing data entry techniques will prove useful. In particular, it will be useful to understand the potential of soft keyboards that could fit on small PDAs and mobile phones. To date, researchers have focused on larger keyboards and simplified tasks. Further, many studies used tablets instead of PDAs and had users employ unrealistic error correction strategies. Understanding the potential of smaller keyboards requires either validated theoretical models or empirical data. This study provides the first empirical evaluation of soft keyboards that are (a) small enough to fit on a PDA or mobile phone, (b) implemented on a PDA, and (c) used to complete realistic tasks.

Second, although Fitts's Law would predict the data entry results observed in this study, it has not been validated in the context of small, stylus-activated, soft keyboards. Further, existing research raises questions about the appropriateness of predictions based entirely on Fitts's Law (e.g., James & Reischel, 2001; Sears et al., 1993). More important, the detailed analysis of the data gathered in this study highlights several results that cannot be predicted using Fitts's Law (e.g., longer times for keys proceeding and following keyboard transitions). Although Fitts's Law may predict the overall results, these predictions are based on an incomplete model of what is occurring when users are interacting with these keyboards. The proposed KLM-style model more accurately describes what occurs when users interact with soft keyboards like those used in this study. As a result, the model can be used to provide an initial assessment of alternative keyboard layouts before resources are committed to implementing and formally evaluating alternative soft keyboard designs.

Third, it is critical to obtain an understanding of data entry rates, error rates, and user preferences when assessing the efficacy of keyboards such as those employed in this study. This suggests that previous studies and existing models are not sufficient. Existing models do not address error rates or user preferences and results from previous studies have suggested that reducing the keyboard size would negatively effect both data entry rates and user preferences (e.g., Sears et al., 1993). Of interest in this study, neither data entry rates nor user preferences were affected by keyboard size. As a result, it may be concluded that soft keyboards can be used effectively on devices in which screen space is severely limited.

The task in which the users were engaged did affect data entry rates. Data entry rates ranged from as low as 7.2 wpm to as high as 12.5 wpm with tasks that required users to switch between the two keyboards resulting in slower data entry rates. This was addressed previously by modeling keyboard transitions. It also appears that tasks that involved larger quantities of text resulted in faster data entry rates. This is addressed by two components of the model presented following: the start-up time required to enter the first character as well as the recovery time required after switching from one keyboard to the other. These results are important because different devices may be used to support different types of activities. Perhaps the most important observation is that switching between keyboards is associated with a significant decrease in data entry rates. By understanding the activities users will engage in and the character sets associated with these activities, keyboards may be designed more effectively such that keyboard transitions can be minimized.

8. CONCLUSIONS

Novice performance with many information technologies can be important for two reasons. First, many individuals make decisions regarding the adoption of various technologies after minimal exposure. If an input technique is difficult when it is first encountered, potential users may choose not to use the device. Second, many information technologies are only used occasionally. Although an input technique that allows rapid data entry after 10 hr of extensive practice may be good from a theoretical perspective, many individuals will never reach this level of expertise. For these reasons, this study focused on novice performance. In this context, the results suggest that soft keyboards that are small enough to fit on mobile phones are not only feasible but that they result in data entry rates, error rates, and preference ratings that are comparable to larger soft keyboards currently used on PDAs. Depending on the task, data entry rates ranging from approximately 7 to 12 wpm were observed.

A detailed analysis revealed that additional time was required when entering the first character of a task. Additional time was also required whenever a user needed to switch between two keyboards. Based on these observations, it is clear that Fitts's Law cannot accurately describe user interactions with soft keyboards like those included in this study. As a result, a KLM-style model was developed that accurately predicts the total time necessary to enter a predefined string of characters. This model may prove useful for evaluating alternative keyboard designs before committing resources for implementation and formal evaluations. This study demonstrated this potential by predicting data entry rates for a variety of data entry tasks if a single keyboard design were used. These predictions highlight potential benefits for tasks involving keyboard transitions but also confirm the lack of benefit for tasks without such transitions. This suggests that a single keyboard design, which eliminates keyboard transitions, may be more effective.

Although confirming that small soft keyboards are feasible, the data also raise questions about the generalizability of the results reported elsewhere. Other researchers have reported data entry rates ranging from 11 to 28 wpm for initial performance and as high as 40 wpm after extensive practice. Given the data entry rates observed in this study, it is critical to consider the experimental methodologies as well as the tasks utilized in each study. Many earlier studies have used paper mockups of the keyboards or a desktop tablet (e.g., Lewis et al., 1999; MacKenzie et al., 1994, 1999). Several studies have had users ignore errors or reenter the correct letter without deleting the incorrect letter (e.g., MacKenzie & Zhang, 1999, 2001). Others have used simple tasks that are not representative of those that users would actually encounter (e.g., MaKenzie & Zhang, 2001; MacKenzie et al., 1999). Many studies have used simplified keyboards with no numbers, punctuation marks, or shift keys, whereas others have used a single keyboard to present all of the available characters (e.g., MacKenzie et al., 1994; MacKenzie & Zhang, 1999). In contrast, participants in this study corrected errors while completing realistic tasks using a Palm-style, two-keyboard design on a PDA.

It is interesting that Lewis (1999) reported faster data entry rates when using a larger, single-panel keyboard on a PDA. Because this study found that keyboard size was not important and suggested that the time required to switch between keyboards is important, it may prove useful to investigate smaller, one-panel keyboard designs in future studies. Note also that the average age of the participants in this study was approximately 26 years. Given the visual and motor demands involved in interacting with small devices and the trend for both visual and motor capabilities to diminish with age, it would be interesting to repeat this research with older participants.

The results suggest several other directions for future research. Keyboard designs that minimize the need to switch between alternative keyboards should be investigated. Keyboards that use smaller keys could provide access to more characters at one time, but readability may become a significant concern. This study employed square alphabetic keys, but future studies may investigate other key shapes. For example, given the narrow displays typically available on mobile phones, it may prove useful to investigate the effectiveness of keys that are taller than they are wide. Finally, designs that automatically provide access to keys users are likely to use may prove useful. For example, a standard QWERTY keyboard could be provided with an additional row of keys that are customized based on the text the user is currently entering or the task the user is performing. Future research could also investigate the integration of models that describe physical movements and visual scanning patterns into the proposed KLM-style model for task completion times, as well as changes that occur as novices gain experience.

REFERENCES

Card, S. K., Moran, T. P., & Newell, A. (1980). The keystroke-level model for user performance time with interactive systems. *Communications of the ACM, 23,* 396–410.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology, 47,* 381–391.

Gentner, D. R. (1983). Keystroke timing in transcription typing. In W. E. Cooper (Ed.), *Cognitive aspects of skilled typing* (pp. 95–120). New York: Springer-Verlag.
Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, *4*, 11–26.
Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, *45*, 188–196.
James, C. L., & Reischel, K. M. (2001). Text input for mobile devices: Comparing model prediction to actual performance. In *Proceedings of the ACM Conference on Human Factors in Computing Systems CHI 2001* (pp. 365–371). New York: ACM.
Lewis, J. R. (1999). Input rates and user preference for three small-screen input methods: Standard keyboard, predictive keyboard, and handwriting. In *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting* (pp. 425–428). Santa Monica, CA: Human Factors and Ergonomics Society.
Lewis, J. R., LaLomia, M. J., & Kennedy, P. J. (1999). Evaluation of typing key layouts for stylus input. In *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting* (pp. 420–424). Santa Monica, CA: Human Factors Society.
MacKenzie, I. S., Nonnecke, B., McQueen, C., Riddersma, S., & Meltz, M. (1994). A comparison of three methods of character entry on pen-based computers. In *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting* (pp. 330–334). Santa Monica, CA: Human Factors Society.
MacKenzie, I. S., Nonnecke, B., Riddersma, S., McQueen, C., & Meltz, M. (1994). Alphanumeric entry on pen-based computers. *International Journal of Human-Computer Studies*, 41, 775–792.
MacKenzie, I. S., & Soukoreff, R. W. (2002). Text entry for mobile computing: Models and methods, theory and practice. *Human–Computer Interaction, 17*,147–198.
MacKenzie, I. S., & Zhang, S. X. (1999) The design and evaluation of a high-performance soft keyboard. In *Proceedings of the ACM Conference on Human Factors in Computing Systems—CHI '99* (pp. 25–31). New York: ACM.
MacKenzie, I. S., & Zhang, S. X. (2001). An empirical investigation of the novice experience with soft keyboards. *Behaviour & Information Technology, 20,* 411–418.
MacKenzie, I. S., Zhang, S. X., & Soukoreff, R. W. (1999). Text entry using soft keyboards. *Behaviour & Information Technology, 18,* 235–244.
Potter, R., Weldon, L., & Shneiderman, B. (1988). Improving the accuracy of touch screens: An experimental evaluation of three strategies. In *Proceedings of CHI '88* (pp. 27–32). New York: ACM.
Pratt, F. (1939). *Secret and urgent: The story of codes and ciphers.* Garden City, NY: Blue Ribbon Books.
Rosenthal, R., & Rubin, D. B. (1994). The counternull value of an effect size: A new statistic. *Psychological Science, 5,* 329–334.
Sears, A., & Arora, R. (2002). Data entry for mobile devices: An empirical comparison of novice performance with Jot and Graffiti. *Interacting With Computers, 14*, 413–433.
Sears, A., Jacko, J. A., Chu, J., & Moro, F. (2001). The role of visual search in the design of effective soft keyboards. *Behaviour and Information Technology, 20,* 159–166.
Sears, A., Revis, D., Swatski, J., Crittenden, R., & Shneiderman, B. (1993). Investigating touchscreen typing: The effect of keyboard size on typing speed. *Behaviour and Information Technology, 12,* 17–22.
Silfverberg, M., MacKenzie, I. S., & Korhonen, P. (2000). Predicting text entry speed on mobile phones. In *Proceedings of the ACM Conference on Human Factors in Computing Systems—CHI 2000* (pp. 9–16). New York: ACM.
Soukoreff, R. W., & MacKenzie, I. S. (1995). Theoretical upper and lower bounds on typing speed using a stylus and soft keyboard. *Behaviour & Information Technology, 14,* 370–379.

APPENDIX

Task One:

Enter this address:
John Doe
8374 Maple Dr.
Apt. 36–C
Baltimore, MD 21250
(410) 391–4398
jdoe@gl.umbc.edu

Task Two:

Enter this URL:
www.giraffe837.com

Task Three:

Enter this URL:
www.travelocity.com/vaca23

Task Four:

Enter this appointment description:
Department Meeting

Task Five:

Enter this appointment description:
Meeting with Bob and Sue about annual budget

Task Six:

Enter this reply to an e-mail:
The meeting this Tuesday has been changed to 2 pm. Please notify me if there is a conflict in your schedule. Bring all materials regarding Alpha project with you to this meeting. I will send more details later in the week.