

Ergonomic and usability tests

Contents

<i>Preface</i>	1
<i>UniTap Device</i>	1
<i>Testing procedure</i>	2
<i>Participants</i>	4
<i>Text entry speed</i>	4
<i>NASA Task load index</i>	6
<i>Learning curve</i>	7
<i>Key strokes per character</i>	8
<i>Resume</i>	9
<i>References</i>	9
<i>Annex 1: NASA TLX Rating Scale Definitions</i>	10

Preface

This document reports the results of the ergonomic and usability tests of the prototype device that was built by RL technologies B.V. in order to evaluate the new text entry technology for mobile devices called UniTap (www.unitap.net). One of the primary purposes of these tests was to make comparison between UniTap and the most relevant technologies, namely Fastap^{TMi}, T9^{®ii} and multi-tap, all of which use buttons (or sensors) activated by a fingertip. At the time of testing only the latter two interfaces were available in commercial mobile phones, while UniTap and Fastap were tested in prototype devices.

UniTap Device



The device that was used for evaluating UniTap technology was designed in a typical form-factor of mobile phone (see picture on the left), so that it can be hold by a hand while typing by a thumb. Mini-buttons were produced as a single piece of rubber mat with conductive roves that matched respective electric circuits on the underlying printed circuit board (however, in retrospect we found this particular principle less efficient compared to membrane switches and other sensor technologies as being quite small in size such buttons could fail to close the respective electric circuit appropriately). These small buttons were left quite noticeable for descriptive purposes; though, they might be visually hidden by certain designing techniques in the

ⁱ www.digitwireless.com

ⁱⁱ www.t9.com



final products.

The built-in software being extremely simple just matched the combination of activated buttons with the character table and if the meaningful combination occurs produced the corresponding symbol on the display. It also made a short beep sound upon entering a symbol. However, one of the features that this software did not have, but which was found quite important for testing purposes was a short timeout after accepting a symbol. Lack of such feature together with rubber nature of the button caused entering several identical symbols during the same tap due to finger/button trembling in some cases. Such effect was considered as a typing mistake during testing sessions and, thus, affected the outcome of the tests.

At our disposal we had UniTap devices with 2 different symbol layouts. First layout conceptually followed regular distribution of Latin alphabet among traditional 12 buttons (as shown on the picture above). The other one had symbols placed in alphabetic order while 2 top rows were reserved for digits. The latter device was used for the tests described in this report.

Testing procedure

Testing procedure was setup as closely as possible to the one described in [1]. This paper reported on the results of comparative analysis of Fastap with T9 and multi-tap text entry methods; thus, following the described procedures allowed us to limit our tests to UniTap device and compare the outcome with the figures presented in this report.

The testing procedure was organized in 3 sessions.

First session measured the **initial reaction** of participants to the UniTap keypad. During this session participants were given very brief introduction to the keypad principle (i.e. “just press to the area where the symbol you wish to enter is printed on the surface”) and then they were asked to enter two sentences – one being traditional sentence and the other being numeric (see examples in Table 1). Finally they were asked to fill in the NASA TLX [2] form (task load index), where they subjectively reported required demands (mental, physical and temporal), their effort, performance and frustration while using the device.

Category	Sentence
Traditional for initial reaction	i bought a cell phone
Numeric for initial reaction	033667001
Traditional	i will be home later.
Non-dictionary	I live on Oak road.
Abbreviated	we can talk 2moro
Numeric	039833298

Table 1. Example of the test sentences used for evaluation.



Second session took place right after the first one had been completed and, thus, measured **novice performance**, e.g. performance of users who had no previous experience with the device. During this session participants were asked to enter four sentences, which were traditional, non-dictionary, abbreviated and numeric (see examples in Table 1). Finally, they were asked to fill in another NASA TLX form.

The final session was organized after participants had been given a reasonable amount of time to intensively practice with the UniTap device (usually, it took place after several days or weeks after the previous sessions); thus, it aimed to estimate **expert performance**. The evaluation procedure was identical to the one utilized for novice performance tests and was finalized by filling in another NASA TLX form. Not all of the participants were trained and took part in the expert performance tests.

For each participant we counted number of finger taps (K) they used for entering each sentence as well as the time (t) they spent. Provided that the respective sentence had C characters we then were able to estimate the following characteristics

Description	Formula
Characters per second	$CPS = \frac{C}{t}$
Words per minute	$WPM = CPS \times \frac{60}{5.98}^i$
Keystrokes per character	$KSPC = \frac{K}{C}$

Table 2. Characteristics calculated for each participant entering each sentence.

Participants were instructed to complete each task as quickly as was comfortable for them; however, they were also asked to correct the mistakes (e.g. wrong or double character) if occurred during the test. Such correction was made by utilization of “backspace” command available on the same keypad (tapping the respective symbol of course contributed to the total number of finger taps K). Timing started when the sentence printed on a sheet of paper was shown to each participant and terminated when the final character was entered and there were no mistakes left in it (if any occurred).

Participants

This report is based on the data collected while evaluating performance of 28 participants (about 30% being females). The majority of the participants were between 15 and 30 years old; however, we had several of them older then 35. They were divided into 4 categories depending on how many SMS per week they typically send, which gave us a notion of how

ⁱ Figure 5.98 is considered to be an average number of characters in the world in English language (as suggested in [1]).

they are familiar with standard text entry methods. These categories were defined as follows: (1) not using SMS at all; (2) 0-5 SMS/week; (3) 5-15 SMS/week; and (4) over then 15 SMS per week. Figure 1 presents distribution of the participants among categories. It is worth noting that quite significant portion of them had had no experience with SMS at all.

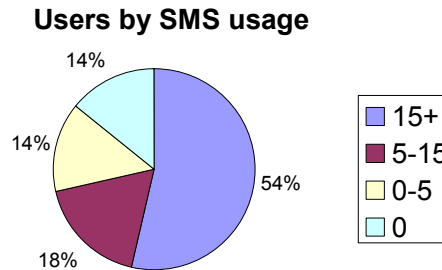


Figure 1. Distribution of the participants with respect to their previous SMS experience (measured in SMS usually sent per week).

It is also important from the evaluation results point of view that none of the participants were native English speakers; however, we had to use sentences in English in order to obtain figures comparable to those presented in [1]. Under such circumstances participants were not likely to memorize the phrase they were given from the first glance (although the sentences were quite simple) and, therefore, had to look at the sheet of paper several times during entering the text. This behavior could significantly affect timing as changing gaze from the device to the sheet and back would take about a second and usually participants had to look at the text 3-4 times per sentence.

About 90% of the participants who reported that they used SMS also reported that they preferred multi-tap rather than T9 (so far, they had not had a chance to compare them to UniTap or Fastap). Such disproportion might, however, have been due to all of the participants being non-native English speakers while some of them might have had not localized mobile phones.

It is also worth noting that more then a half of participants who frequently used SMS reported that they usually typed text with 2 hands and, thus, they did so with UniTap device as well.

Text entry speed

Figure 2 presents the comparative view of average text entry speed during initial reaction session. T9 appeared to perform best for the first task, which was simple traditional sentence, due to previous experience of most participants with T9 (as reported in [1]). Multi-tap and T9 performed poorly for numeric sentence as the entry mode was initially set to "text", which was quite inefficient for entering digits.

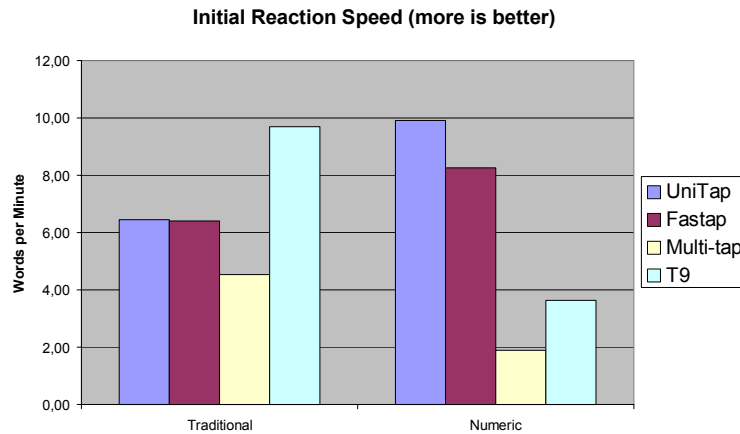


Figure 2. This chart shows average speed for initial reaction session.

Figure 3 demonstrates average speed of text entry achieved during second session for four different types of sentences. One of the reasons of Fastap’s better performance for numeric sentence might be our use of alphabetic symbol layout with digits put together in 2 rows (see UniTap Device section), while Fastap utilizes traditional telephone style digit placement which most of the participants had got used to.

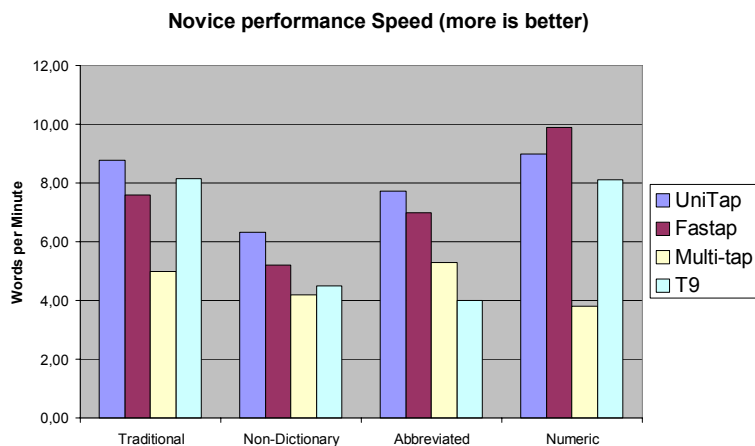


Figure 3. This chart shows average speed for novice performance session.

Figure 4 shows average speed of text entry for expert participant. It was quite expected that well-trained T9 users would demonstrate very good performance for traditional sentences consisting of dictionary words.

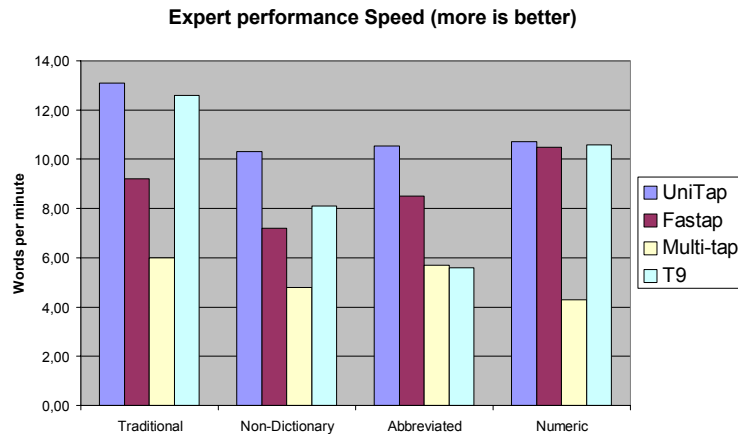


Figure 4. This chart shows average speed for expert performance session.

NASA Task load index

The following three figures (Figure 5, Figure 6, and Figure 7) present NASA Task Load Index [2]. This index is a multi-dimensional rating procedure that provides an overall workload score based on six subscales: Mental Demands, Physical Demands, Temporal Demands, Own Performance, Effort, and Frustration level (see details in Annex 1: NASA TLX Rating Scale Definitions). Although it is possible to make a single index by weighted averaging of these parameters, we preferred considering them separately following the procedure described in [1].

These figures demonstrate that UniTap device performed better with respect to all parameters except level of frustration. Our participants reported that the current design of the keypad revealing many small buttons on a surface looked a bit frustrating since common practice of button usage was pushing each one at a time. However, such appearance was designed for descriptive purposes; though, final product might look quite differently as discussed in UniTap Device section.

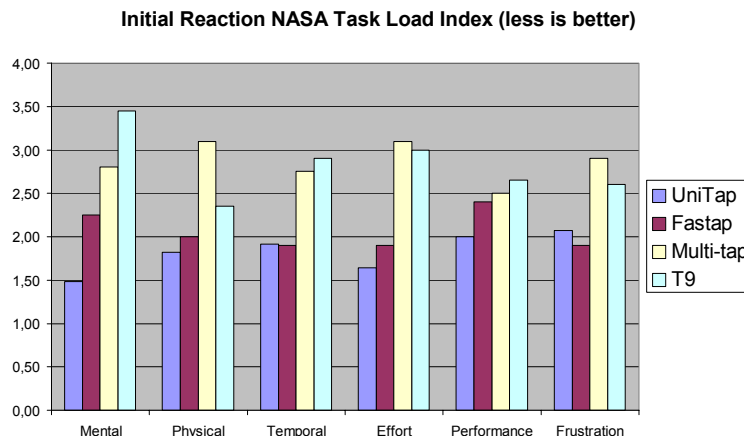


Figure 5. This chart shows average task load index for initial reaction session.

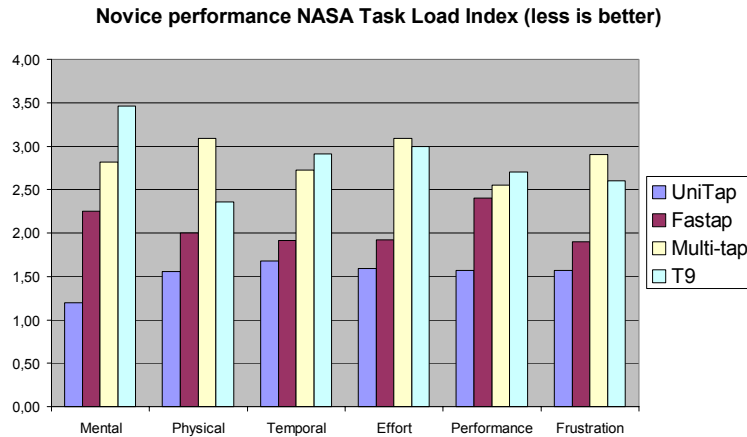


Figure 6. This chart shows average task load index for novice performance session.

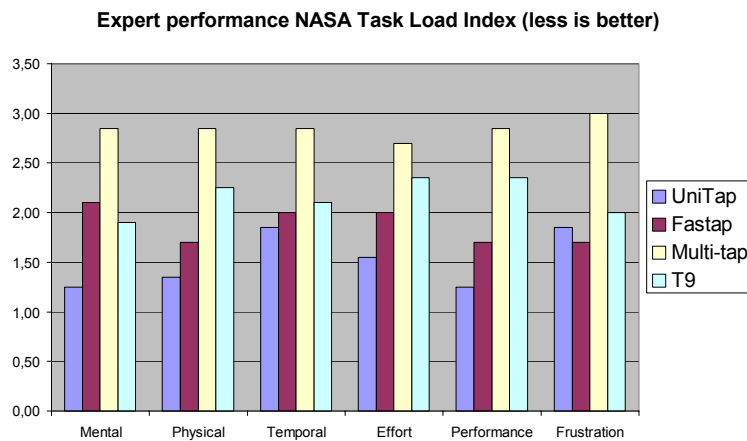


Figure 7. This chart shows average task load index for expert performance session.

Learning curve

Charts in this section (Figure 8) present learning curves for different classes of sentences and different text entry methods. They show that T9 successfully compete with UniTap and Fastap on traditional sentences where all words are available in dictionary. Expert users of T9 also managed to enter numeric sentences quite fast as they probably learnt how to switch it to numeric mode. For other types of sentences UniTap demonstrated best performance on the average and quite stable learning curves.

Figure 9 shows aggregate learning curve obtained by averaging learning curves for respective sentences. It demonstrates that UniTap interface performed better for initial reaction as well as novice and expert users.

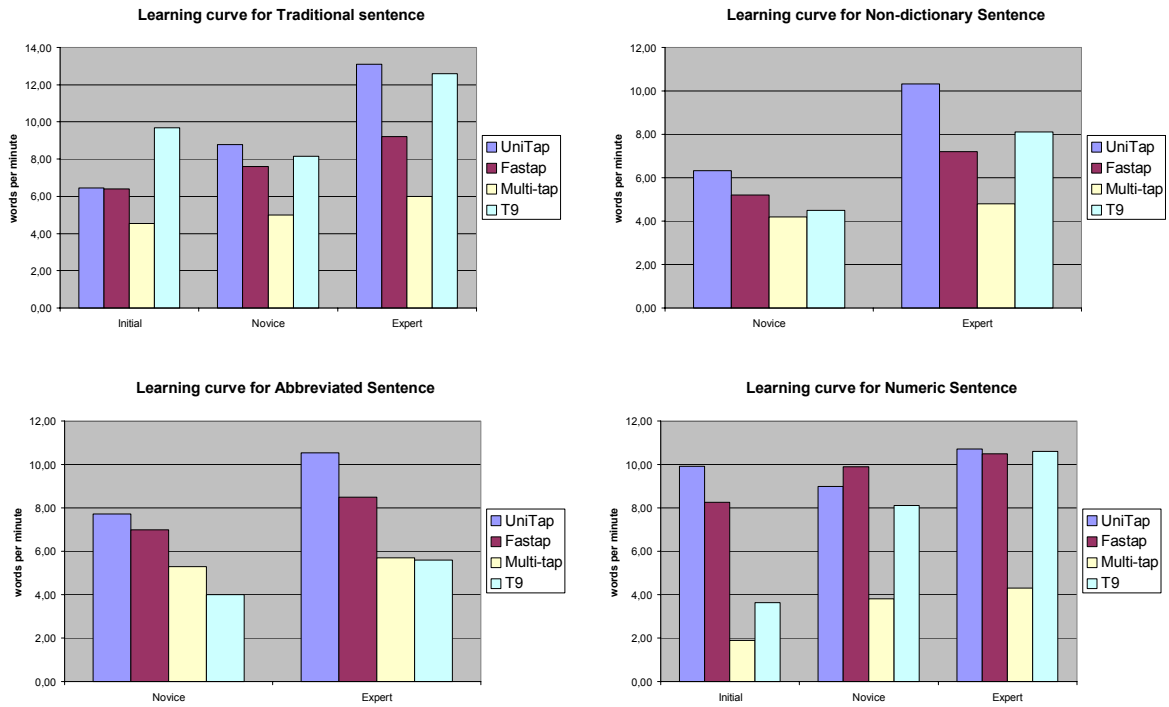


Figure 8. These charts demonstrate learning curves for different classes of sentences.

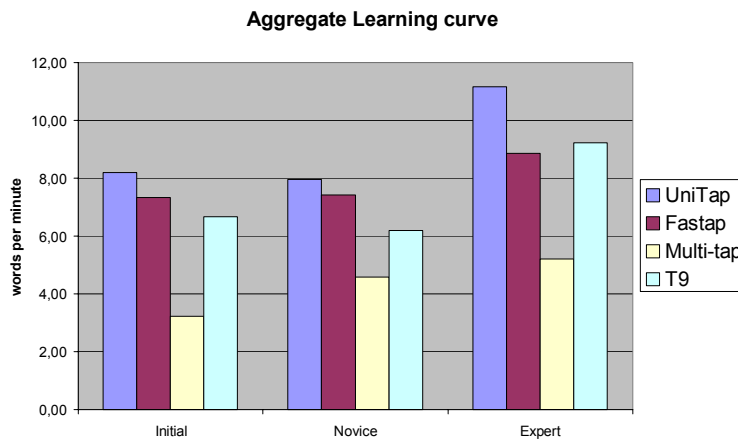


Figure 9. This chart demonstrates an aggregate learning curve.

Key strokes per character

KSPC parameter was measured in order to evaluate efficiency of each interface with respect to a number of finger strokes required for entering a symbol. It also reflects number of errors being made while using respective interfaces because participants were asked to correct each error they make.

Figure 10 presents average number of key strokes made for entering one character during different evaluation sessions. It demonstrates quite stable performance of UniTap and

Fastap interfaces, while KSPC for Multi-tap and T9 noticeably decreases as participants are being trained. It is worth noting that Fastap and Multi-tap KSPC for expert users is larger compared to the respective figures of novice ones. According to [1], it might have happened due to participants' trading a decrease in accuracy for an increase in speed; however, UniTap does not demonstrate such effect, which suggests that speed and accuracy for this interface do not correlate much for measured level of performance.

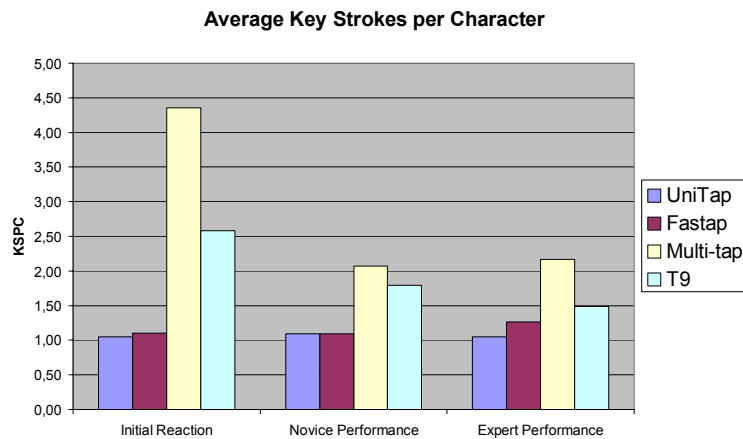


Figure 10. This chart shows average number of key strokes made for entering one character during different evaluation sessions.

Resume

This report presented the results of the comparative evaluation of different text entry interfaces namely UniTap, Fastap, Multi-tap and T9. Obtained results demonstrate that UniTap interface performs better than other technologies that have been tested in most cases and, which is more important, its performance is comparatively stable for different types of sentences and all categories of users (novice as well as expert). UniTap principle has also been reported less demanding on the average with respect to workload, which was measured by NASA Task Load Index technique.

It was also noted that prototype device that was used for testing had some technical issues which, however, would be resolved in the final products. Therefore, commercial products featuring UniTap technology are likely to perform even better.

References

- [1] Amal Sirisena. *Mobile Text Entry*. Report made under supervision of Andy Cockburn at the Department of Computer Science, University of Canterbury, Christchurch, New Zealand, November 8, 2002.
- [2] S. Hart, L. Staveland. Development of NASA TLX (Task Load Index): Results of Empirical and Theoretical Research. P. Hancock and N. Meshkati eds., *Human Mental Workload*, Elsevier Science, pp. 139-183.

Annex 1: NASA TLX Rating Scale Definitions

Title	Endpoints	Description
Mental demand	Low/High	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
Physical demand	Low/High	How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
Temporal demand	Low/High	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Effort	Low/High	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Performance	Good/Poor	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
Frustration level	Low/High	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?