JUS
JOURNAL OF USABILITY STUDIES

Issue 2, Vol. 1, February 2006, pp. 91-108

Empirical Evaluation of a Popular Cellular Phone's Menu System: Theory Meets Practice

Sheng-Cheng Huang
School of Information
The University of Texas at Austin
1 University Station D7000
Austin, TX 78712-0390
Email: huangsc@mail.utexas.edu

I-Fan Chou
School of Information
The University of Texas at Austin
Email: ifan@mail.utexas.edu

Randolph G. Bias
School of Information
The University of Texas at Austin
Email: rbias@ischool.utexas.edu

Abstract
A usability assessment entailing a paper prototype was conducted to examine menu selection theories on a small screen device by determining the effectiveness, efficiency, and user satisfaction of a popular cellular phone's menu system. Outcomes of this study suggest that users prefer a less extensive menu structure on a small screen device. The investigation also covered factors of category classification and item labeling influencing user performance in menu selection. Research findings suggest that proper modifications in these areas could significantly enhance the system's usability and demonstrate the validity of paper-prototyping which is capable of detecting significant differences in usability measures among various model designs.

Keywords
Empirical findings, laboratory study, menu selection, model-based evaluation, paper prototyping, usability data analysis, usability method.

Introduction

With the march of technology, today's cellular phone interfaces attempt to provide much visual interaction

with a user via a limited-size display on a handset, the screen of the typical mobile phone being relatively tiny compared to a regular computer monitor. For this reason, the interface design of a menu system on a cellular phone's screen of 160x160 or 240x160 pixel-per-inch resolution has analogous concerns that engineers encountered on an early 80s' computer output screen.

Many of the early studies of menu selection design on computers focused on the cognitive factors of a menu's hierarchical structure and the structure's impact on end users' behaviors and performance in information retrieving. Since Miller (1981) raised the question, the tradeoffs of a menu's breadth and depth has been debated, considering factors such as visual search time, time of motor/machine response, and the limitations of human working memory. Except for Billingsley's (1982) proposal of using a map aid, findings related to this issue consistently suggested the advantage of employing a broader menu structure to achieve better user performance and accuracy (Allen, 1983; Burns et al., 1986; Kiger, 1984; Norman, 1991; Parush and Yuviler-Gavish, 2004; Seppala and Salvendy, 1985; Tullis, 1985). However, arguments in favor of a broad menu structure fall flat when considering a device with a small screen. With less space to display information, designers of cell phones and personal digital assistants (PDAs) tend to chunk menu items of a broader menu into several pages or screens. Therefore, end-users must employ more scrolling operations and increase working memory load in searching and navigation, which at the same time reduces the speed and accuracy in the use of the menus (Parush and Yuviler-Gavish, 2004; Tang, 2001; Ziefle, 2002). However, these studies tend to be based on the presumption of the perfect classified and categorized menu items in which the menu structure was the only independent variable affecting end-users' performance outcome. This setting excluded other possible influential factors such as the relevance of the menu items' classification and their labeling. Despite being aware that "choices of menu options often involve inferences about logical and/or categorical relationships among items" (Allen, 1983), few early studies practically considered the influences of the classification and the labeling. Although some argued for the most appropriate number of a menu's categories in each level (Kiger, 1984; Norman and Chin, 1988; Seppala and Salvendy, 1985), none reflected that ill-classified items or ill-labeled categories could be the causes of a menu's breadth/depth tradeoffs in user performance.

Norman and Chin (1988) noted that the number of categories and labels were altered simultaneously when the menu's depth or breadth was changed, which brought additional uncertainties when the choices between categories increased. The increase of categories not only affected the load of classification, but also required designers to assign new labels to them. Meanwhile, labels assigned to categories or menu items had tradeoffs such as providing longer (specific) descriptions or shorter (partial) ones, which as well affected end-users' search processes. A more specific and descriptive naming was more helpful to users to reduce the uncertainty of the menu item's definition, but it occupied more screen space and fewer items could be displayed at a time (Norman, 1983). Considering the small physical display of a mobile phone, it is more likely that labels serve as partial descriptions of the menu's items, which makes naming the alternatives more difficult. Time spent by the user

in selecting among alternatives affects the efficiency of menu retrieval (Lee and MacGregor, 1985). Consequently, in order to achieve clear distinctiveness between alternatives of menu items, assigned terms must emphasize the differences and avoid illustrating the commonality of their functions (Norman, 1991, p.142). Moreover, the recent study of Ziefle (2002) identified that even when a logical term was assigned to a menu item, users could be misled when their expectations of the item did not match the term. Due to these constraints, appropriate and careful naming is critical to reduce ambiguity in a menu selection system, especially on small-screen cell phones.

Another question surrounding menu selection is "Can the design ease a user's workload by reducing the task complexity"? Salvendy and Jacko (1996) provided evidentiary support of Campbell's (1988) four characteristics of a complex task: (1) having multiple paths to complete a task, (2) having multiple outcomes, (3) having conflicts in selecting paths to achieve different outcomes in a task, and (4) having uncertainty in selecting potential paths. In order to create a comprehensive and meaningful menu, Shneiderman (1998) suggested that distinctive categories should be formed by task-related objects and actions and the organization should appear relevant to users' tasks. As a result, it is safe to say that cell phone users, like all computer users, expect to see a menu organized to fit the tasks they perform. In this study, we proposed a different and broader perspective on menu selection design of a small screen device. Instead of addressing only the breadth/depth tradeoffs of menu structure, our interests extended to two interdependent factors: the category's classification and the menu item's labeling, as we tested the influences of these two factors on users' task performances. Further, we intended to demonstrate the viability of paper prototyping as a test method for such small-screen menu structures.

Research Objective and Questions

The subject we were interested in testing was the menu of Nokia ™'s "Series 40 Developer Platform 1.0," which is the most popular platform on Nokia ™'s devices. The motivation of examining it was that we felt that the current Series 40's menu structure was flawed in classification and labeling of its menu items. Thus, one objective was to generate a hypothetically improved prototype based on multiple menu design theories so that we could test it to see if there were significant improvements in user performance. User performance was evaluated by the usability measures that ISO 9241-11 suggests: (1) effectiveness of the system; (2) efficiency of the system; and (3) user satisfaction. In our study, effectiveness was determined by the success rate, efficiency was determined by time and attempts spent on individual tasks, and the level of user satisfaction was measured by the participants' survey data and interview comments. Two specific questions to be addressed included:

1. Do menu category classification and menu item labeling have significantly interactive impacts on user performance of a small screen device's menu selection?
2. Can we demonstrate the efficacy of a usability evaluation that we design for a paper prototyping variation that is relative to Snyder and other practitioners have suggested?

Methods

There were three separate components of our study. First, we presented an open, online survey to gather some demographic data and to solicit feedback on cell

phone use. Second, we performed a pilot study on an existing cell phone menu to identify potential usability problems with the structure. Third, after having redesigned the menu based on theories of menu design and our pilot study findings, we tested a subset of the respondents to the online survey, comparing their performance on the extant menu structure and on our redesign.

Participants

The participants in this study were selected from the aforementioned 41 subjects responding to the open online survey held at the University of Texas at Austin from June 1 through June 18, 2004. Nineteen participants were selected for the comparison test, and assigned to one of two groups. We generated matched groups that had approximately equal numbers of each gender and of varying experience with a Nokia ™ cellular phone. Table 1 shows the general characteristics of the grouped participants.

Table 1. Characteristics of Test Participants

	Group 1 (n=9)	Group 2 (n=10)	Total (n=19)
Male	5	5	10
Female	4	5	9
Nokia ™ user	5	6	11
Non-Nokia ™ user	4	4	8

Type of Assessment

Nielsen (2003) and Snyder (2003) have recommended the method of paper prototyping in usability design because of its ability to bring to bear end-user data on design concepts early in the product development cycle. The usability comparison test we demonstrated here was based on a variation of a paper prototype assessment that is relative to Snyder and other practitioners have suggested. Since we did not have the ability to make direct modifications on a mobile phone's menu system, paper prototyping was a method, whereby we could compare the old and the new menu structures. In addition, it is important to note that this study focused only on the information organization of the menu system. Therefore, not implementing an actual handset test eliminated the impacts of other factors that might affect the participants' performance, such as interaction design of the panel and screen layout.

Test Materials

Two menu models were employed in the evaluation. Model A and Model B were paper replicas of two menu systems. Model A was the original menu of Nokia ™ 6610 cellular phone, and Model B was the prototype menu developed by the investigators. Model B was generated to obtain evidentiary support of the modification impacts on structure, classification, and labeling. Both models were represented in the form of a map of the menu hierarchy in paper for three reasons: (1) to exclude factors of a cell pone's operational interface that might affect a user's task performance; (2) to avoid procedure errors of test administrators presenting incorrect menu items or level structure during the tests that could happen in a traditional method of using index cards, and (3) to increase the opportunity for the participants to explore the whole menu structure, which offers the most possible menu items that can be found in a cell phone. Modifications in Model B included:

Structure modification: to reduce the breadth and depth of the menu's structure, which corresponded to

the observation of previous studies made on small screen devices (Parush and Yuviler-Gavish, 2004; Tang, 2001; Ziefle, 2002):

1. The number of menu items on the first level was reduced from 13 to 8 by moving some isolated functions ("Alarm clock", "Radio", "Gallery", and "Organizer") on the first level to the category of "Extra" and the repositioning of the category of "Application" (also see modification 4).
2. The breadth of "Settings" category was reduced by taking off the category of "Tone settings" (also see modification 5).

Classification modification: to perform better grouping in menu categories, which was based on user comments given in the pilot test of this study:

3. The function of "Welcome note" was moved from "Phone settings" to "Display settings". "Phone settings" was renamed (also see modification 7).
4. The category of "Application" was moved from the first level to the category of "Service". The names of these two categories were changed (also see modification 7).
5. Functions of "Tone settings" and "Profile" were combined into one category named "Tone setting profiles" which replaced the category of "Profile".

Labeling modification: to eliminate ambiguous naming of menu items, which applied Norman and Shneiderman's naming principles and Ziefle's observation result (Norman, 1991, p.142; Shneiderman, 1998; Ziefle, 2002):

6. Labels regarding the functions of web services were renamed: "Service" became "Web service," "Home" became "Web browser," and "Service inbox" was renamed to "E-mail notice."
7. Several other changes on labels were made: "Connectivity" became "Device connectivity," "Phone setting" was renamed to "Device setting," and "Applications" became "Java applications."

Tasks

Tasks were designed to provide scenarios that mimicked real-life situations of using a cell phone. For example, to motivate the participant to check received calls, we offered the scenario: "You are talking to friend A on your cell phone and friend B cuts in. Unfortunately, you have to hang up on A but you promise to call A back later. You finally finish talking to B and want to contact with A again. You need to check the calls you've received recently in order to get A's number in the last call and call A back." Five representative tasks were chosen based on the online survey of participants' most frequent use of handset operations such as "phonebook inquiry" and "setting options." In addition, tasks were to (1) be representative of the whole population of tasks, and (2) capture some of the major modification differences between Model A and Model B. The five tasks were:

1. Check received calls.
2. Find the wireless Internet access.
3. Find the option, "Welcome Note".
4. Turn on vibrating alert.
5. Set the phone on the silent mode.

The tasks were to be completed in the sequence provided above. The sequence was determined by the task complexity suggested by Campbell's (1988) definition of "multiple paths" and "uncertainty." Having the largest number of multiple paths to achieve the goal and alternatives of choices among menu items in Model A, task 5 was regarded the most complex task by the investigators.

Procedure

The usability evaluation was run in 19 individual sessions. Each session included three portions completed by the participants in the following order of a repeated measures design to counterbalance the order in which the two models were seen:

1. Each participant in Group 1 finished a performance evaluation in which a series of five tasks was completed on Model A. The same performance evaluations in Group 2 were completed on Model B.
2. Each participant in Group 1 finished a performance evaluation in which the same tasks were completed on Model B. The same performance evaluations in Group 2 were completed on Model A.
3. After the evaluation, every participant in Group 1 and Group 2 was interviewed to gather additional insights regarding users' subjective perceptions of these two models.

Task evaluation of the paper model was based on path-finding behavior that simulated the scroll-and-click actions that users apply with the actual mobile phone's interface. This is achieved by using an index card having a cut-out window to represent the screen of a cell phone and placing it over the paper replica of the menu model. The participant will move this card along the paths; see the menu items through the cut-out window, and select the options with a highlighter (see figure 1). This paper-based simulation allowed us to reveal only one level of each menu structure and a few menu items at a time to the participant, which also reduces the chance for the participant to learn the items in the lower or upper levels of each menu. The final result of the participant's activities will be a map of traces in the menu system (see figure 2).

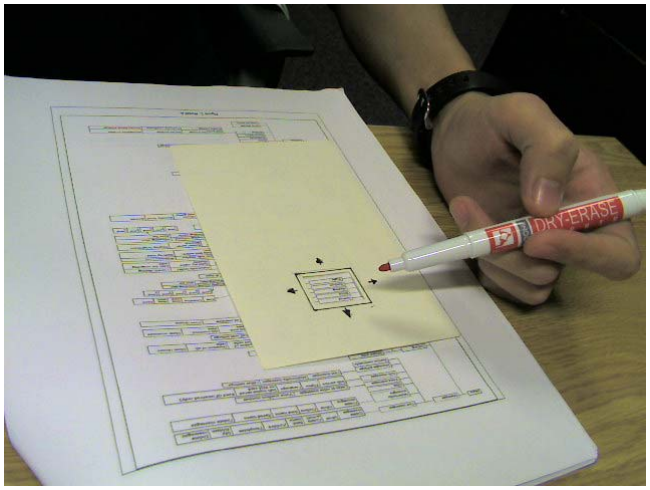

Figure 1. Paper Simulation of Menu Selection on Cell Phones

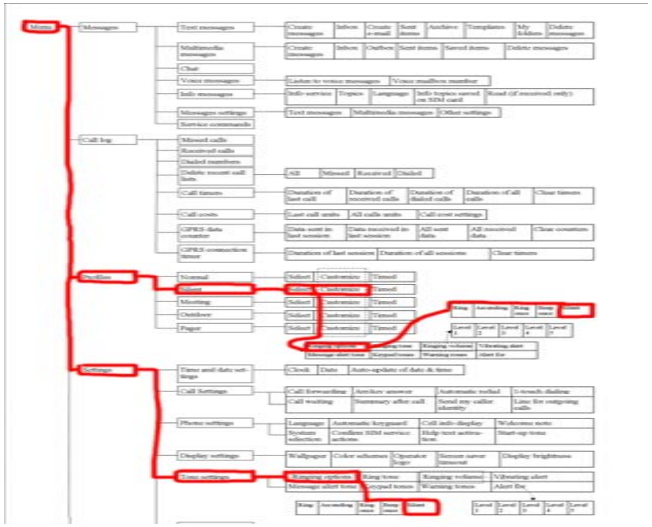

Figure 2. Example of a Participant's Traces of Menu Selection

The evaluations were timed, video taped, and observed by the test administrators. The participants were encouraged to complete the tasks without guidance and verbalize their thoughts during the process. The investigators requested the participants declare if they had completed the task successfully, and did not inform them if they had completed it per expected outcomes. The participants were allowed to withdraw from the evaluation if they felt frustrated or for any reason. (None chose to.) We pre-set an upper limit of five attempts per task to prevent the over-extending our allocated time, however, test participants were not informed of this limit.

Measures

Measures included:

1. The time to complete each task.
2. The number of attempts to complete each task.
3. Task success rate.
4. Number of and types of errors:
   Observations and Comments: evaluation monitor notes regarding when participants had difficulty, when an unusual behavior occurred, or when a cause of error became obvious.
   Non-critical Error: a participant made a mistake but was able to recover during the task in the allotted attempts.
   Critical Error: a participant made a mistake and was unable to recover and complete the task successfully. The participant might or might not realize that a mistake had been made.

Results

Time, number of attempts, and success rate were analyzed quantitatively to identify whether there was any significant difference in test participants' performance between Model A and Model B. We concentrated also on identifying the type of errors and illustrated participants' reasons of making mistakes and attitudes toward the menus in order to understand users' satisfaction level. The results also included participants' suggestions and recommendations for possible improvements of the menu system.

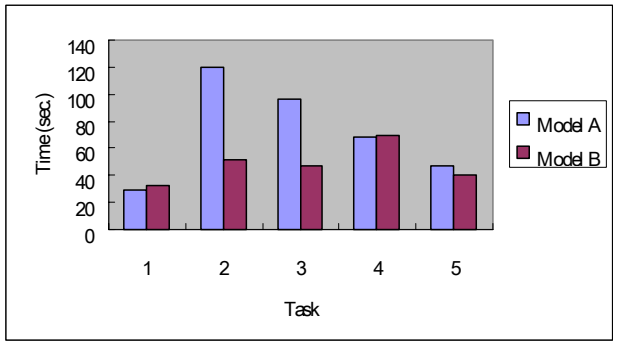

Figure 3. Average Time Span for Tasks in Model A and B

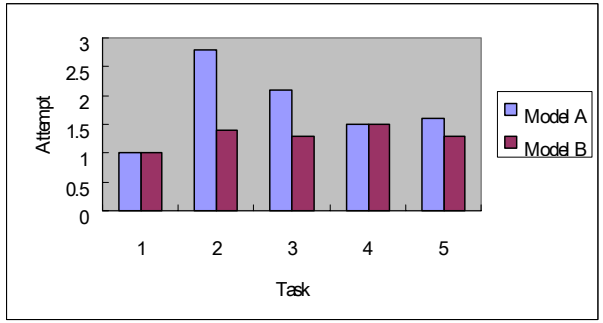

Figure 4. Average Attempts for Tasks in Model A and B

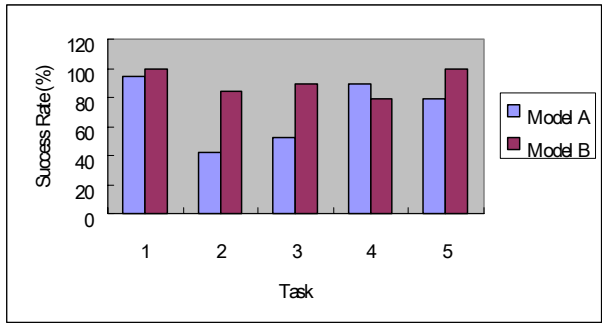

Figure 5. Success Rates for Tasks in Model A and B

Table 2. Tests of Within-Subjects Contrasts that Yielded Significant Differences

Measures	Variance ratios and alpha levels
Task 2 Time Span between Model A and Model B	F (1, 17) = 19.655, p < 0.001
Task 2 Attempts between Model A and Model B	F (1, 17) = 16.436, p = 0.001
Task 3 Time Span between Model A and Model B	F (1, 17) = 10.959, p = 0.004
Task 3 Attempts between Model A and Model B	F (1, 17) = 5.799, p = 0.028
Total Task Time Span between Model A and Model B	F (1, 17) = 20.947, p < 0.001
Total Task Attempts between Model A and Model B	F (1, 17) = 21.706, p < 0.001
Task 2 Time Span between Model A and Model B with Co-variance of being a Nokia ™ user or not	F (1, 16) = 6.372, p = 0.023
Task 2 Attempts between Model A and Model B with Co-variance of being a Nokia ™ user or not	F (1, 16) = 5.510, p = 0.032

The results indicate that the participants (1) spent less time to complete three of the five tasks in Model B, averaging a total of 119.4 sec less time across all five tasks (a 33% improvement); (2) had fewer attempts to complete the tasks in Model B (6.6 compared with 9.1, across all five tasks); and (3) had higher success rate in completing the tasks in Model B (90.5% compared to 71.6% across all five tasks). Figure 3 and Figure 4 illustrate that the participants spent less time and fewer attempts on Model B to complete the tasks, especially in Tasks 2 and 3. Although people tended to take longer and make more attempts on tougher tasks, neither Task 2 nor Task 3 was the most complex task among the tasks according to Campbell's (1988) definition. Thus, there was no speed-accuracy tradeoff but other factors were involved to produce this outcome. On the other hand, as seen in Figure 5, the participants completed tasks with lower success rate in Model A, especially in task 2 and task 3. A two-factor repeated measures ANOVA was applied to the raw data of each task. Three significant performance differences in time and attempts were found in Task 2, Task 3, and Total task accumulation between the two models. Considering the co-variance of being a Nokia ™ user or not, there was also a significant performance difference in time and attempts in Task 2, suggesting that being a Nokia ™ user did not account for the user performance difference. Although the analyses show that only Task 2 and Task 3 showed statistically significant differences in performance, the overall performance between Model A and Model B was as well significant, indicating that the participants performed better in Model B than Model A regardless which model they evaluated first. Table 2 summarizes the results of ANOVA tests.

In order to address the factors causing the performance differences in Model A and Model B, the errors made by the participants in each task were analyzed with human error identification (HEI) techniques and the method of hierarchical task analysis (HTA), which allowed us to break a task down into a number of operations after the initiation (Jordan, Thomas, and Weerdmeester, et al., 1996). Tables 3 to 7 present the path of attempts on which non-critical errors occurred and compare the paths made by the participants in both models in each

task. Non-critical errors were separated into two types, actual errors and uncertain errors. Actual errors were defined as the error attempts different from the correct path. Uncertain errors were defined as the attempts in which the participants followed the correct path, but they did not continue finishing the correct path because they weren't confident they were right.

Task 1: Check received calls

With the smallest number of trials and the highest success rate among all five tasks, the participants only made one particular error in selecting alternatives under the correct category in both models in task 1. This may explain the failure for any Model A vs. Model B difference to obtain; we experienced a "floor effect" in this frequent task.

Task 2: Find the wireless Internet Access

In both model A and B, the participants need only three correct selections to complete task 2, which means that the path complexity is at the same level in both models. The investigators found that the reduction of the breadth of the menu's first level alone was not the major factor which helped the participants perform better in Model B (Model B has only eight selective items instead of 13), which indicates there were other factors affecting the participants' performance. Table 3 shows that the participants made more error attempts in Model A than in Model B, which caused significantly worse performance in time span in Model A (Table 2). For Model A, most of the errors in Task 2 were made in the attempts of selecting the correct category on the first level. The participants wandered between alternatives such as "Connectivity," "Extras," "Applications," and "Services." Only one participant looked for Internet access, in this case, homepage of the service provider, under "Services" in the first attempt. In addition, 16 (84%) participants assumed that "Connectivity" was a more relevant option to connect to the Internet after considering all the alternatives on the first menu level. Moreover, the participants also assumed that the Internet access was under "Applications" and "Extras." The participants commented that terms such as "Services," "Applications," "Connectivity," and "Extra" were not specific enough to lead them to determine the definitions of the subcategories. Hence, the participants would select one of these options, apparently at random, and keep trying to learn from mistakes they made. Not only did these labels appear unclear and incomprehensive, but also some terms were too technical to be understood by the participants. For instance, "Infrared" and "GPRS" were two terms that the participants were unclear about.

Although almost half of the participants (47%) considered that "Service" category might relate to the Internet access, after selecting and browsing the menu items, they did not think any of the subordinate options would connect them to the Internet. For example, "Home," the name of the function allowing users to connect to the service provider's homepage, confused the participants in this particular case. Seven participants commented that they supposed that "Home" was the service provider's customer service call. The term, "Home," appeared difficult to associate with an online homepage or the Web connection. "Service inbox" was also not descriptive according to the participants. The participants could not associate "Service inbox" with E-mail checking, which was the actual function of the option. Therefore, even the participants who happened to select the correct path or

option still were unsure about their choice because of the terms offered in Model A. These errors indicate that ambiguity in labeling in Model A misled the participants to select wrong options and caused a low success rate (42%).

Table 3. Task 2 Error Attempts

Correct Path: **Menu→Services→Home (A)** **Menu→Web services→Web browser (B)**						
	Group 1+ Group 2, Model A			Group 1+ Group 2, Model B		
	Error attempts	N	%	Error attempts	N	%
Actual Errors	Menu→ Connectivity→(GPRS)	16	84.21%	Menu→ Device connectivity	3	15.79%
	Menu→ Applications	7	36.84%	Menu→ Web services→ Go to address	2	10.53%
	Menu→Extras	4	21.05%	Menu→ Web services→ Bookmarks	2	10.53%
	Menu→Settings	3	15.79%	Menu→Extras	1	5.26%
	Menu→Services→ Go to address	2	10.53%			
	Menu→Call log	1	5.26%			
	Menu→Organizer	1	5.26%			
	Menu→Settings→ Phone settings	1	5.26%			
Uncertain Errors	Menu→Services	9	47.37%			

On the other hand, the participants made fewer errors and had higher success rate (84%) in Model B. The participants commented that they expected to see terms like "Web" offered by Model B in the option's labeling to represent wireless Internet access instead of just having "Services" in Model A. They also commented that making selections between "Web services" and "Device connectivity" in Model B was more distinctive than "Services" and "Connectivity" in Model A. The results supported Norman's (1983) and Shneiderman's (1998) suggestion of using specific and distinctive labeling. In addition, the participants had much better performance in Model B in time and attempts (Table 2). Even those participants who were already Nokia ™ users still performed significantly less well with Model A in time and attempts (Table 2). These significant differences suggest that our grouping and labeling modifications in Model B had positive impacts on user performance.

Task 3: Find the option, "Welcome Note"

Again, Task 3's path complexity is the same in both models (three correct selections), whereas the participants made more error attempts in Model A than in Model B (Table 4). Most errors made in Task 3 were caused by selecting alternative subordinate categories under "Settings." This pattern could be seen in both models. "Settings" in Model A stands along with 48 menu items separated in eight subordinate categories. Not only is classification difficult with such a large number of options, but making the subordinate categories distinctive from each other also requires careful labeling. Regarding the classification problem, in Model A, more than 94% of the participants assumed that "Welcome note" was an option underneath "Display settings" rather than "Phone settings." Moreover, the other two options under "Display settings," "Operator logo" and "Wallpaper" also misled the participants. Even though they could not find the matching term of the option, they would rather guess similar alternatives under "Display settings" than go back to the upper level to retry. The second highest error attempt in Model A is "Profile" because the

participants stated that they regarded "Welcome note" as a personalized setting. On the contrary, in Model B, where "Welcome note" is under "Display settings," the error attempts were much fewer than Model A. Moreover, the participants' performances are significantly better in Model B in time and attempts (Table 2). The result designates that shifting "Welcome note" to "Display settings" appears to better match the classification intuitive to participants.

Table 4. Task 3 Error Attempts

Correct Path: **Menu→Settings→Phone settings→Welcome note (A)** **Menu→Settings→Display settings→Welcome note (B)**						
	Group 1+ Group 2, Model A			Group 1+ Group 2, Model B		
	Error attempts	N	%	Error attempts	N	%
Actual Errors	Menu→Settings→ Display settings	18	94.74%	Menu→Settings→ Device settings	5	26.32%
	Menu→profiles	7	36.84%	Menu→Messages→ Info messages	2	10.53%
	Menu→Settings→ Display settings→Operator logo	5	26.32%	Menu→Settings→ Device settings	1	5.26%
	Menu→Settings→ Display settings→Wallpaper	3	15.79%	Menu→Settings→ Device settings→ Cell info display	1	5.26%
	Menu→Extras	2	10.53%			
	Menu→Settings→ Accessory settings	2	10.53%			
	Menu→Services→ Settings	1	5.26%			

Task 4: Turn on vibrating alert

Since Model A has two outcomes to successfully achieve task 4, the path is presumably considered more complex than the one in Model B. Based on the observations during Task 4, in Model A, the participants had difficulties distinguishing the differences between "Phone settings," "Call Settings," and "Tone settings" under "Settings." The results are shown in Table 5.

Most of the participants' incorrect attempts happened because of the indistinctiveness among these subordinate categories. The likelihood of the incorrect attempt occurrences in "Phone settings," "Call settings," and "Tone settings" are 10 ~ 20% for each participant, across both Models. Therefore, the labeling of these subordinate categories in Model A appeared to be less effective in distinctiveness, which made the participants unable to intuitively determine the correct selection between categories. Because of the enormous number of sub-categories in "Settings," the menu structure became four levels deep, which required that the participants to spend more time navigating to the bottom level where the target options located. One participant complained about the overly expanded depth of "Settings" and expressed obvious frustration. The participants mentioned that they were having trouble visualizing the whole structure of the menu system. On the other hand, despite the fact that Model B reduced the task complexity of multiple paths by combining Model A's options of "Profile (first level category)" and "Tone setting (sub-category in 'Settings')" together, the participants' performance was not significantly better. The cause was associated with the structure problem in depth and the management of identical menu items. The target option was on the deepest level of the menu and the very same option also repeatedly appeared in each forth-level selective items of a third-level subordinate category. Model A has the above-mentioned structure problem in "Profile" as Model B's "Tone setting profile." Thus, the single modification of Model B's structure in this particular task was not effective in enhancing the user

performance. Further reduction of menu structure and better categorization of identical menu items are anticipated to improve this issue.

Table 5. Task 4 Error Attempts

Correct Path: **Menu→Settings→Tone settings→Vibrating alert (A)**
Menu→Profiles→Silent*→Customize→Vibrating alert (A)
Menu→Tone setting profiles→Silent*→Customize→Vibrating alert (B)

	Group 1+ Group 2, Model A			Group 1+ Group 2, Model B		
	Error attempts	N	%	Error attempts	N	%
Actual Errors	Menu→Settings→ Phone settings	4	21.05%	Menu→Settings→ Call settings	3	15.79%
	Menu→Settings→ Call settings	2	10.53%	Menu→Settings→ Device settings	3	15.79%
	Menu→Profiles→ Silent→Select	2	10.53%	Menu→ Tone setting profiles→ Silent→Select	3	15.79%
	Menu→Settings→ Tone settings→ Ringing options	2	10.53%	Menu→Settings	1	5.26%
	Menu→Settings→ Tone settings→ Ringing volume	1	5.26%	Menu→ Tone setting profiles→ silent→customize→ Ringing options	1	5.26%
	Menu→Profiles→Silent*→ Customize→ Ringing options	1	5.26%			
Uncertain Errors				Menu→ Tone setting profiles	2	10.53%
				Menu→ Tone setting profiles→ silent→customize	1	5.26%

Task 5: Set the phone on the silent mode

Task 5 is a related task to Task 4, where the path complexity remains more difficult in Model A than the one in Model B. Although Model A's "Profile" is supposed to serve as a "shortcut" for the user to select ringing options more quickly, Table 6 indicates the ambiguity of "Profile" and "Settings" in Model A regarding their associations with the ringing tones. First, the label "Profile" does not correctly represent its item's functions according to the participants. Eight participants (42%) presumed "Profile" was either user profile or personal profile, whereas the defined description of the "Profile" category resembles "custom ringing tones in different environment settings." Failing to be offered an appropriate label, in Model A, only one participant selected "Profile" in the first attempt. Instead, they proceeded to search the options under the category of "Settings," where more expanded options exist. Nine participants (47%) ignored the "Profile" and moved directly to "Settings." This indicated that the participants rarely associated "Profile" with "Ringing tone settings" or "Environment mode," instead, they regarded "Profile" as personalized settings or user information. The reason that the participants hesitated to select "Profile" in this task was not only the irrelevant labeling, but also the unfamiliarity of its options.

On the other hand, Model B offers only ringing tone settings in the "Tone setting profile" category and put the "Silent" option on a closer-to-the-surface level (the second level of the menu). Although Model B's modification reduced the types of error attempts, this change did not significantly improve the participants' speed or success rate. (It must be remembered that all test participants conducted the tasks in the same order. It could be that the relatively quick speed and low error rates in the later tasks were due to the participants

having stumbled upon the correct menu item during earlier tasks.)

Table 6. Task 5 Error Attempts

Correct Path: **Menu→Profiles→Silent →Select (A)** **Menu→Settings→Tone settings→Ringing options→Silent (A)** **Menu→Tone setting profiles→Silent→Select (B)** **Menu→Tone setting profiles→Silent*→Customize→Ringing options→Silent (B)**						
	Group 1+ Group 2, Model A			Group 1+ Group 2, Model B		
	Error attempts	N	%	Error attempts	N	%
Actual Errors	Menu→Settings→ Tone settings→ Ringing volume	4	21.05%	Menu→Settings	2	10.53%
	Menu→Settings→ Call settings	3	15.79%			
	Menu→Settings→ Phone settings	2	10.53%			
	Menu→Profiles→Meeting→ Select	2	10.53%			
	Menu→Profiles→Silent*→ Customize→Ringing volume	1	5.26%			
Uncertain Errors	Menu→Profiles→Silent*→ Customize	1	5.26%	Menu→ Tone setting profiles→ silent→customize	4	21.05%

Despite the fact that 10 participants used the options under "Profile" in Model A, seven of these 10 participants (70%) did not complete the task correctly with the "Select" option, which also identifies another labeling problem. One participant commented that the "Customize" options under each environment setting were redundant because they repeatedly offered five groups of identical menu items, which already existed in "Tone settings" under the "Settings" category. Also, in Model A, there were four error attempts (21%) in "Tone settings" under the "Settings" category.

Regardless of the fact that the correct path was under the "Tone settings" category, the participants had difficulty making selections among the subordinate categories. The participants often wandered between "Ringing options" and "Ringing volume." Four participants (21%) assumed "Silent" mode should be classified under "Ringing volume" rather than "Ringing options." The distinctiveness between these two categories is low. In addition, the term, "Ringing volume" suggests that users can turn off the sound of the phone; however, this category has no option allowing the users to turn the volume to "0" (silent). Thus, although the mutual connection between "Silent" mode and "Ringing volume" was identified, the system's effectiveness was weakened for not offering relevant functions matching users' needs.

Interview Data

Table 7. Average Score in a Scale of 1 (worst) ~ 5 (best)

Model A	Group 1 (n=9)	Group 2 (n=10)
Structure	3	2.5
Classification	3	3
Labeling	2.5	3

Table 8. Preference between Model A and B (n=19)

Model A (original)	Model B (mockup)
0%	100%

Satisfaction: Table 7 summarizes the results of the participant's average rating of Model A. The scores were given regarding the breadth and depth of the level (structure), the grouping of menu items (classification), and the naming of the categories and options (labeling). Although both groups offered similar and marginally satisfactory scores regarding the menu's

structure, classification and labeling, all participants favored Model B's structure (Table 8). The participants' main criticism of the structure of Nokia ™ 6610 menu system (Model A) is that the first level has too many orphaned options. For example, single functions such as "Alarm clock" and "Radio" stand alone instead of being categorized. The Nokia-experienced participants reflected that some of the options such as "Gallery" and "Organizer" on the first level were rarely used in their experience. They recommended that the menu's first level should be reduced by integrating these orphaned functions such as "Games," "Gallery," "Organizer," "Alarm clock," and "Radio" together under one category or under the "Extras" category like the mockup (Model B) managed. The participants proposed that in order to avoid an overly expanded breadth, only the most frequently utilized functions should be placed on the first level of the menu, which included "Messages," "Settings," and "Call logs."

In the interview section, participants were asked to point out confusing labels (terms used in naming the categories and options) in Model A. Results are shown in table 9. "Profile" was heavily criticized by the participants for not descriptively representing "the custom ringing tones in different environment settings." Beside the most confusing label, "Profile" (100%), discussed in the Task 5's error analysis, the second most confusing labels are all related to wireless Internet services. In table 9, labeling terms with asterisks such as "Service", "Connectivity", "GPRS," "Application," and "Home" were categories and options that the participants would make attempts to complete Task 2. Having struggled in this particular task, the participants complained that these labels were either not descriptive or specific enough for them to determine the function's definition at the first glance. Some participants stated that they would try these options on a real handset to see the outcomes in order to learn their actual functions. However, the participants as well concurred that these labels could be more explicit so that they could have completed the task without spending much time on trial-and-error. These findings indicate that Model A's labeling design was not intuitive enough to enhance the efficiency of the system. On the other hand, the participants were impressed by the much more specific labeling offered by Model B.

Table 9. Confusing Labels in Model A (n=19)

	N	%
Profile	19	100%
Service*	13	68.4%
Connectivity*	10	52.6%
GPRS*	8	42.1%
Application*	8	42.1%
Extra	7	36.8%
Home*	5	26.3%
Gallery	5	26.3%
Setting	2	10.5%
Ringing Option	2	10.5%
Customize under "Profile"	2	10.5%
Operator Logo	2	10.5%
Info. Message	2	10.5%
Message	1	5.3%
Go to Address	1	5.3%
Inbox	1	5.3%
Chat	1	5.3%

Conclusion

The expansion of additional options makes the menu system more complex in both breadth and depth and

increases memory load of users to learn the system (Allen, 1983 and Billingsley, 1982). Having this particular obstacle, the participants commented that they might have performed more poorly on a real handset evaluation with this breath and depth of the original menu. This supported the existence of the breadth/depth tradeoffs in menu design identified in the early studies (Allen, 1983; Kiger, 1984; Norman, 1991; Parush and Yuviler-Gavish, 2004; Seppala and Salvendy, 1985; Tullis, 1985). Although the studies of breadth/depth tradeoffs proposed that broader menus had better performance, outcomes of this study suggest an opposite user preference of a less extensive menu structure on a small screen device. This result supported the suggestion of not having a broad menu structure on a small screen (Parush and Yuviler-Gavish, 2004; Tang, 2001; Ziefle, 2002). The over-expanded breadth and depth of the menu requires the users to navigate and select options they do not need, which often results in the additional consumption of time and reduced system efficiency (Lee and MacGregor, 1985), not to mention increased human memory load. Due to the constraint of the limited display, the effect of breadth/depth tradeoffs in navigation is much more obvious and important in a small screen device. Being asked to improve this issue on a real handset, the participants preferred to have a more visibility of menu items. To achieve this, we suggest: (1) reduce both breadth and depth of the menu and (2) instead of displaying only a limited few items on one screen, put more menu items and options in one page so that users can avoid extra scrolling actions on a level.

These two objectives can be achieved by performing a better classification on menu items and by minimizing the redundant placement of menu items. Results of this study indicated that a weak classification of a menu system is injurious to system usability. Despite the fact that experienced Nokia users and inexperienced users could successfully navigate through the menu, irrelevantly organized menu items significantly reduced user performance. Evidence found in Task 3 suggested that end-users' perspectives could be quite different from the designer's presumption. The efficiency of participants' menu selection within this system was often constrained by the ill-labeled categories (Task 2). There was an obvious disparity between the manufacturer's manner of defining the options' names and how the users comprehend the meaning of the labeling. For example, the naming of the "Profile" category was the most confusing section of the whole menu (table 14), poorly representing the function definition and the participants' expectations of the term. This supports Ziefle's (2002) finding that the term to be assigned not only should appear logical to the menu item, but also should match users' expectation. Hence, in order to avoid potential misleading, the labeling design must consider the names close to users' mental models which match the task scenarios. Of course, the application of a user-centered design approach (e.g., Vredenburg, Isensee, and Righi, 2002) to menu structure is the best way to accomplish this. In addition, specific and descriptive terms were often required by the users to help them determine the correct selection and distinguish the differences between menu items, which supported (1) Norman's (1983) suggestion of using longer (and more complete) descriptions in naming is more useful to users learning the system; and (2) Shneiderman's (1998) recommendation of forming distinctive menu categories based on users' tasks.

Even though Model A was rated satisfactory by the test subjects, minor flaws of that original menu in fact accumulated and increased the chance of obstacles for the users to operate it. The distribution of time and attempts spent in learning to complete a single operation was quite large. The participants spent 197 to 666 seconds and 6 to 13 attempts to complete all five tasks in Model A. In contrast, with minor adjustments in the areas we proposed, user performance was significantly improved both in time and attempts. The participants spent only 115 to 370 seconds and made only 5 to 8 attempts to complete all five tasks in Model B. Not only were the system's overall effectiveness and efficiency enhanced, the elevation of user satisfaction was also exhibited by the participants' universal preference of the new prototype. Findings of this study revealed that ill-categorized and ill-labeled menu items had strong impacts on user performance in menu selection (Task 2 and Task 3). Moreover, the uncertainty increased by these two factors as well increased the task complexity in choosing the correct path or item among multiple alternatives. Our study demonstrated that our modifications were applicable and effective in enhancing user performance. It is also important to note that the results offered support for the efficacy of a paper prototyping variation as a way of testing the usability of an information architecture, which shows that an evaluation of a user's path-finding activities in a map of menu hierarchy can detect significant differences in user performance among a variety of models. In the future, we plan to keep interests in further assessing the aspects of a small-screen device's menu design and developing new methods for usability evaluations in paper prototyping.

Practioner's Take Away

- The effect of breadth/depth tradeoffs in navigation is much more obvious and important in a small screen device. Findings suggest (1) reduce both breadth and depth of the menu and (2) display more menu items and options in one page so that users can avoid extra scrolling actions on a level.
- Ill-categorized and ill-labeled menu items have strong impacts on user performance in menu selection. Findings support (1) Norman's (1983) suggestion of using longer (and more complete) descriptions in naming is more useful to users learning the system; and (2) Shneiderman's (1998) recommendation of forming distinctive menu categories based on users' tasks.
- This study demonstrates the efficacy of a paper prototyping variation as a way of testing the usability of an information architecture, which shows that an evaluation of a user's path-finding activities in a map of menu hierarchy can detect significant differences in user performance among a variety of models.

Acknowledgments

We would like to express our appreciation to Dr. Kay Lewis, Ms. Lisa Kleinman, Dr. Andrew Dillon, Dr. Avi Parush and all the participants who contributed their assistance in this study and the manuscript of the paper.

References

Allen, R. B., (1983). Cognitive factors in the use of menus and trees: an experiment. *IEEE Journal on Selected Areas in Communications*, 2, 333-336.

Billingsley, P. A. (1982). Navigation through hierarchical menu structures: does it help to have a map? *Human Factors Society, 26th Annual*

Meeting (Santa Monica, CA: Human Factors Society), pp. 103-107.

Burns, M. J., Warren, D. L. and Rudisill M. (1986). Formatting space-related displays to optimize expert and nonexpert user performance. ACM SIGCHI Bulletin, *Proceedings of the SIGCHI conference on Human factors in computing systems CHI '86*, 17(4).

Campbell, D. J. (1988). Task complexity: a review and analysis. *Academy of Management Review*, 13, 40-52.

Hornof, A. and Kieras, D. (1997). Cognitive modeling reveals menu search is both random and systematic. In *Proc. CHI '97 Conference: Human Factors in Computer Systems*, pages 107-114. ACM Press. Available at http://www.acm.org/sigchi/chi97/proceedings/paper/ajh.htm

Jordan P. W., Thomas, B. and Weerdmeester, B., et al. (1996). *Usability Evaluation in Industry.* London: Taylor & Francis.

Kiger, J. (1984). The depth/breadth trade-off in the design of menu-driven user interfaces. *International Journal of Man-Machine Studies*, 20:201-213.

Lee, E. and MacGregor, J. (1985). Minimizing user search time in menu retrieval systems. *Human Factors*, 27:157-162.

Miller, D. P. (1981). The depth/breadth tradeoff in hierarchical computer menus. *Proceedings of the 25th Annual Meeting of the Human Factors Society*, 296-300.

Nielsen, J. (2003) *Paper prototyping: getting user data before you code.* Retrieved July 27, 2004 from http://www.useit.com/alertbox/20030414.html .

Norman, D. A. (1983). Design principles for human-computer interaction. *Proceedings of CHI'83*, pp. 1-10.

Norman, K. (1991). *The Psychology of Menu Selection: Designing Cognitive Control of the Human/Computer Interface*. Norwood NJ: Ablex Publishing Corporation.

Norman, K. and Chin, J. (1988). The effect of tree structure on search in a hierarchical menu selection system. *Behaviour and Information Technology*, 7(1):51-65.

Parush, A. and Yuviler-Gavish, N. (2004). Web navigation structures in cellular phones: the depth/breadth trade-off issue. *International Journal of Human-Computer Studies*, 60, 753-770.

Salvendy, G. and Jacko, J. (1996). Hierarchical menu design: Breadth, depth, and task complexity. *Perceptual and Motor Skills*, 82:1187-1201.

Seppala, P. and Salvendy, G. (1985). Impact of depth of menu hierarchy on performance effectiveness in a supervisory task: Computerized flexible manufacturing system. *Human Factors*, 27:713-722.

Shneiderman, B. (1998). *Designing the User Interface: Strategies for Effective Human-Computer-Interaction.* Reading, MA: Addison-Wesley.

Snyder, C. (2003). *Paper Prototyping: The Fast and Easy Way to Design and Refine User Interfaces.* San Francisco CA: Morgan Kaufmann Publishers.

Tang, K. E. (2001). Menu design with visual momentum for compact smart products. Human Factors, 43(2), 267-277.

Tullis, T. S. (1985). Designing a menu-based interface to an operating system. *Proceedings of CHI'85*, pp. 79-84.

Vredenburg, K., Isensee, S., and Righi, C. (2002). *User-Centered Design: An Integrated Approach.* Upper Saddle River, NJ: Prentice Hall PTR.

Ziefle, M. (2002). The influence of user expertise and phone complexity on performance, ease of use and learnability of different mobile phones. *Behaviour and Information Technology*, 21(5), 303-31

Sheng-Cheng Huang is a doctoral student in the School of Information at the University of Texas at Austin and works as a graduate research assistant of the research project studying the effects of Microsoft ClearType™ on end-users under the supervision of Randolph G. Bias, Ph.D. His main interests focus on usability of human-computer interaction (HCI) and semiotics in the computer-mediated communication (CMC) environment.

I-Fan Chou is a MSIS graduate of the School of Information at The University of Texas at Austin, where she works as a teaching assistant of Usability Engineer, Advanced Usability, and Research Method. She did her Master's project on interface usability evaluation of a proof-of-concept handheld device at IBM Pervasive Computing Lab in Austin, Texas. She currently works as a user experience and interactive design intern at Punchcut, focusing mainly on mobile interface design and mobile applications from user experience perspective.

Randolph G. Bias worked in industry for over 20 years as a usability engineer, helping software developers make human-computer interfaces user friendly. In 2003 he came to the School of Information to research human information processing and human-computer interaction. Randolph has written over 50 technical articles in the area of human information processing, and co-edited *Cost-Justifying Usability, 2nd edition: An update for the information age* (R. G. Bias and D. J. Mayhew, Eds., 2005, San Francisco: Morgan Kaufmann).