Research Article

Evaluation of a Navigation Radio Using the Think-Aloud Method

Paul A. Green¹ and Jin-Seop Park²

¹ *Driver Interface Group, University of Michigan Transportation Research Institute, 2901 Baxter Road, Ann Arbor, MI 48109-2150, USA*

² *Advanced Vehicle Safety and Dynamics Research Office, Korea Automobile Testing and Research Institute, 200 Samjon-ro, Songsan-myeon, Hwaseong-si, Gyeonggi-do 445-871, Republic of Korea*

Correspondence should be addressed to Paul A. Green; pagreen@umich.edu

Received 10 October 2012; Accepted 26 December 2012

Academic Editor: Motoyuki Akamatsu

Copyright © 2013 P. A. Green and J.-S. Park. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In this experiment, 13 licensed drivers performed 20 tasks with a prototype navigation radio. Subjects completed such tasks as entering a street address, selecting a preset radio station, and tuning to an XM station while “thinking aloud” to identify problems with operating the prototype interface. Overall, subjects identified 64 unique problems with the interface; 17 specific problems were encountered by more than half of the subjects. Problems are related to inconsistent music interfaces, limitations to destination entry methods, icons that were not understood, the lack of functional grouping, and similar looking buttons and displays, among others. An important project focus was getting the findings to the developers quickly. Having a scribe to code interactions in real time helped as well as directed observations of test sessions by representatives of the developers. Other researchers are encouraged to use this method to examine automotive interfaces as a complement to traditional usability testing.

1. Introduction

People want products that are easy to use, and that is particularly true of motor vehicles. Numerous methods have been developed to assess the ease of use of driver interfaces, both traditionally, and more recently from the human-computer interaction literature [1–3]. The three most prominent methods are (1) usability testing [4–8], (2) expert reviews [9–11], and (3) the think-aloud method [12–15]. Methods vary in terms of their value for formative evaluation (while development is in progress) and summative evaluation (at the end of development). See [16] for an extensive overview of how various methods are conducted and where they should be applied.

Usability testing is the gold standard of usability test methods, as it involves real users performing real tasks, though often in a laboratory setting, and can be part of either formative or summative testing. The purpose is to determine task completion times and errors. Generally, usability testing occurs in the latter stages of design, when a fully functioning interface is available. Usability tests are time-consuming to plan and analyze and can be costly.

Consequently, there has been considerable interest in predicting user performance, in particular task time [17–22]. Task times for experienced users can be predicted in a fraction of the time to plan, conduct, and analyze a usability test. If the method used by subjects to perform a task is known, the predictions should be as accurate as the usability test data [23].

Expert reviews can be an efficient alternative to usability testing, especially early in design, though they may be used for summative testing as well. In an expert review, each step of each task is examined to determine how the interface should be designed according to established usability heuristics and guidelines. Expert reviews are often criticized as being “just someone’s opinion.” Therefore, reviewers should be professionally certified in human factors or usability. (See <http://www.bcpe.org/>).

In the think-aloud method, users describe their logic as they try to use an interface. For example, a subject might say, “I selected the city name but cannot figure out how to get to the next step,” or “Sometimes, there is an OK button in the lower right corner, but there is not one here.

FIGURE 1: Most common screen.

FIGURE 2: One of two destination screens.

FIGURE 3: XM radio screen.

FIGURE 4: List of previous destinations screen.

I am stuck and frustrated.” The think-aloud method helps evaluators identify what is confusing or misleading and how those problems can be resolved. If subjects fall silent, then the experimenter prompts them to speak but needs to do without interfering with the subject’s thinking process or influencing them. When to prompt and what to say is much more difficult to do than it may seem. In fact, considerable experimenter training and practice are required. See [16] for a discussion.

Think-aloud evaluations are most useful during the early stages of design while the design is still being formulated, identifying problems users experience more readily than other methods. Unfortunately, as is described later, data reduction in think-aloud evaluations is very time-consuming.

Recently, the authors conducted an evaluation of an early prototype of the Mobis Generation 3 navigation radio for Hyundai-Kia vehicles [24]. The complete experiment included four parts (1) a think-aloud evaluation from a human factors expert, (2) a think-aloud evaluation involving 13 ordinary drivers, (3) a follow-up survey of those drivers primarily concerning their understanding of icons, and (4) estimates of task times per SAE Recommended Practice J2365 [20]. Because this experiment was conducted during the early stage of interface design, and the interface designers needed to know what problems users would encounter, the focus of the experiment and this paper concerns only part 2, the think-aloud evaluation by ordinary drivers.

There are many other ways this data could have been collected. For example, questions concerning what subjects did and what subjects did and why, could have been asked retrospectively. In selecting methods to utilize, the authors considered the specifics of the request for quote from the sponsor, verifying conformance to accepted industry practice (e.g., SAE J2365), what information was believed to be most

useful to the sponsor, the experience of the research team, the funding for the project, the schedule, and other factors. There was extreme pressure to complete this project very quickly to meet the production schedule set by Hyundai-Kia. Therefore, considerable thought was given as to how to complete this project quickly, which meant that less time was spent on certain activities than is ideal, and methods to accelerate data collection and analysis were explored.

2. Method

2.1. Navigation Device Examined. The device examined was an early working prototype of a Mobis generation 3 navigation radio. As shown in Figure 1, the navigation radio consisted of an LCD display surrounded by 10 hard buttons (e.g., select satellite radio, seek), two CD related buttons, a volume knob, and a tuning knob. These hard buttons as well as soft buttons on the touch screen allowed access to hundreds of screens. Figures 2, 3, and 4 show example screens.

2.2. Test Facility. The experiment was performed using the third generation UMTRI driving simulator while “parked.” The navigation system was mounted into the center stack of the simulator cab. To enable signal reception and use of the GPS and XM functions, an antenna was installed, connecting the simulator lab with an outside room. Figures 5 and 6 show a hypothetical subject being recorded, the equipment used, and the recorded image from an actual subject. Although the cameras were in plain sight, subjects ignored them, in part because the camera in front of them was small.

2.3. Sequence of Tasks in the Experiment. In each session, one session per subject, subjects (1) completed biographical and consent forms and had their vision checked, (2) practiced the

FIGURE 5: Experimenter and a hypothetical subject interacting. Note: not shown are the Panasonic WJ-420 quad splitter, the Sharp LC10-A3U-B 10 in LCD monitor, the Panasonic DMR-EZ47V DVD recorder, the AudioTechnica microphone, or the Mackie DX8 digital audio mixer.

FIGURE 6: Sample screen grab from a typical subject.

think-aloud method, (3) completed 20 test tasks in a fixed order while thinking aloud, and (4) completed a 19-page survey requested by the sponsor. Practice involved counting the number of chairs in the place where they lived, going room by room. As intended, the subjects did not simply list how many chairs were in each room, but said something about the kind of chairs present, where the chairs were located in each room, or other information. If they did not provide some of those details, then a question was asked to encourage them. Often, the process of recalling chairs was a virtual journey ("The front door takes me into the living room. In that room... Next to that room, around the table ..."). The practice was quick and helped subjects understand what was meant by "think aloud."

Some 20 tasks were examined (Table 1). These tasks were selected because of their importance, frequency of occurrence, and to provide data on a variety of entry types. Only manual entry was allowed. For ease of administration and analysis, the order of tasks was fixed.

Subjects were not given any documentation or instruction as to how the interface functioned, as the interface design was intended to be intuitive.

The second author served as the experimenter for the main experiment. He had reviewed the literature on think-aloud studies and also served as the experimenter for pilot

testing of the first author, which led to discussions of when to prompt and what to say. The second author also observed and provided feedback on test sessions of the first few subjects after each session was complete. Ideally, more time would have been desired for training, but that was not feasible given the sponsor's product development schedule.

During the think-aloud test, the experimenter, seated in the front passenger seat, presented a sheet of paper, one per task, describing the task to perform and the data to use, and asked the subject to think aloud while doing the task. When subjects fell silent for an extended period of time (typically 30 s, but sometimes longer), the experimenter prompted them. ("What are you looking for? What are you expecting to see? Are you confused?") Generally, few prompts were needed. There were no specific rules about which prompt to use when. Although these prompts seem leading, it was often apparent from what subjects said and their actions what that the subject state was consistent with these prompts. For example, if a subject repeatedly switched between screens but did not select anything else, then "What are you looking for" was an appropriate prompt.

If a subject was not making any progress after three to five minutes on a task (depending on the time remaining), they were given a hint. Hints identified what to do next (e.g., press this button), without any explanation of why an action was appropriate. If subjects continued to struggle, they were told to stop and move on to the next task, as the intent was to reveal as many problems as possible in the time available.

A scribe (a very fast typist) sat outside the simulator cab and attempted to record verbatim what the experimenter and subject said during the experiment. The only and very general instructions to the scribe were to record everything said and to use the video and audio recordings to fill in any gaps. In part, a verbatim transcription was feasible because the experimenter and subject were not talking continuously, so the scribe was able to catch up during pauses. Anything that the scribe missed (e.g., if the subject spoke quickly) was filled in immediately after each test session (in the 30 minutes or so before the next subject) from the audio and video recordings. Having a complete transcription essentially immediately after each session was completed shortened the time to reduce the data. Furthermore, after the fact review of the audio recordings, collected using a system that was hastily assembled and not optimized for recording quality, were sometime inaudible, especially when subjects mumbled. Although requiring a scribe in addition to an experimenter made the experiment more difficult to schedule, given when this experiment was conducted, a scribe was always available.

2.4. Subjects. Thirteen licensed drivers volunteered to serve as subjects: 6 younger people (ages 19–26, 3 men and 3 women) and 7 older people (ages 65–83, 3 men and 4 women). They were recruited via an advertisement on Craigslist. All older subjects were retired. Five of the younger subjects were students. Younger subjects were the most likely users of the navigation radio, especially of the audio functions. Older subjects were those most likely to be challenged by the interface and would be the first to encounter problems

TABLE 1: Task list.

Category	#	Task	Data to use
Navigation	1	Enter destination via street address method	100 North 5th Avenue; Ann Arbor, MI
	2	Enter waypoint via POI (Point Of Interest) method	Nearest McDonalds from current position
	3	Enter destination via POI method	Nearest gas station from current position
	4	Cancel the route	Current route
	5	Enter destination via POI method	Detroit Metro Airport
	6	Change the map scale	Zoom in
	7	Enter destination via street address method	1600 Pennsylvania Avenue; Washington, DC
	8	Enter destination via intersection method	US-23 and I-94
	9	Enter destination via POI method	Cobo Hall in Detroit, MI
	10	Enter destination via intersection method	Broadway St. and Plymouth Rd.
	11	Store address	Subject's home address
CD	12	Insert CD and select track	CD provided, third track
FM/AM radio	13	Tune radio	FM 91.7
	14	Adjust volume	To a comfortable level
	15	Set preset	FM 91.7 as the Preset 1
Satellite radio	16	Adjust volume	To a comfortable level
	17	Tune XM radio	Elvis Station
	18	Turn off XM radio	Turn off
Clock	19	Change and set time	11:45 pm
iPod	20	Connect iPod and select song	"Wonderful Tonight"

this study was to reveal. All subjects had corrected visual acuity adequate to drive. They drove a mean of 8,000 miles per year, somewhat less than is typical in the United States.

Other than being a licensed driver, in good health, in specific age categories, a native English speaker, and experience with XM/Sirius radio, there were no other requirements to participate. Thus, there was no control over experience with technology in this experiment, a limitation not included so subjects could be recruited in the time frame available. Experience with relevant technology was mixed. Five subjects owned an iPod, and one owned another brand of MP3 player. Only two of the subjects had vehicles with XM/Sirius radios. Five subjects owned GPS systems: 3 TomToms and 2 Garmins.

Subjects were paid \$40 for their time if they completed the experiment in two hours. Subjects who took longer (some of the older subjects) were paid an extra \$10.

The sample, seemingly small, was more than adequate for identifying problems, of which 64 were identified. "A problem was defined as a situation where a task took too long, subjects struggled to make progress, or they otherwise expressed doubt ("I am not sure what button to press"), confusion, irritation ("I would like to shoot the engineer that designed this"), or other undesired feelings" ([24] page 48). Problems invariably involved deviations from the intended sequence of steps to complete the task expeditiously. Research shows that after about six subjects or so, the number of new problems found with each additional subject is small, with the specific number varying the problem severity and

other factors [25, 26]. Specifically, the six subject value comes from assuming that each subject has about a one in three chance of discovering a problem, that problems are independent, and the goal is to discover 90% of the problems. Furthermore, this initial analysis assumes all problems are of a similar severity, and one may have different goals for different level of severity. (See [27] for the most recent of a long series of papers, reports, and now a book chapter on sample size.)

Although one can quibble on the specifics of the calculations, the surprise to many nonhuman factors, nonusability professionals is that most of the problems can be found with just a few more than a handful of properly selected subjects. The data from this experiment confirmed that conclusion (Figure 7), with most problems being found by the first few subjects. Testing more subjects would provide better statistical evidence for the frequency of occurrence of each problem and would identify more problems, but not many. Most importantly, testing more subjects would delay producing a complete report informing the designers of interface problems to be corrected. In this case, boosting the confidence of the sponsor was partly why more than six subjects or so were tested. Further, in these instances, when deciding which aspects of an interface to modify, the percentage of subjects who encounter a problem may be secondary. Rather, if just one subject encounters a problem, and the problem seems reasonable, then changes to eliminate or reduce the impact of that problem should be considered.

FIGURE 7: Cumulative usability problems found as a function of number of subjects. Note: this figure determines the % of problems (out of 64 total) found utilizing data from subjects in the order they were tested. So, combining the data of subjects 1 and 2, $38/64 = 59\%$ of the problems were found.

3. Results

3.1. Data Reduction for Problem Identification. Data reduction consists of (1) listening to each session to verify each transcription, correcting them as needed, (2) identifying each problem that users experienced, and (3) identifying the frequency, severity, and persistence of each problem. A problem was indicated when a task took too long, subjects struggled to make progress, or they otherwise expressed doubt (“I am not sure what button to press”), confusion, or irritation (“I would like to shoot the engineer that designed this”).

Subjects indicated problems in several ways. Indications included (1) questions (“What is this? Where is the button? How do I get the map?”), (2) statements of uncertainty (“I’m not sure if I can have the map when I click on this button. It does not allow me to save this radio frequency. I am wondering why it’s not accepting this.”), and (3) exclamations with filler words (“Oh, man...,” “Umm, ...”). About 13% of the problems for younger subjects were associated with questions, whereas for older subjects, questions were linked to a third of the problems. In contrast, about 81% of the problems for younger subjects were associated with statements of uncertainty, versus 61% of the problems for older subjects. There was no difference in the use of filler words (about 7% for both age groups). Specific examples of how particular problems identified appear later in this paper.

There were instances during the experiment where the subject was silent for an extended period of time, where there were no probes from the experimenter, and where it was uncertain from the transcript and video recording what the subject was thinking. This often occurs with novice experimenters as they focus on observing what subjects do. One solution would be a timing device prompted the experiment (unobtrusively to the subject) to probe the subject to think aloud.

Frequency is the number of times the problem occurred, usually across subjects, but sometimes within subject groups. Persistence can be the number of times a problem reoccurred within each subject.

There were three primary severity categories—critical (“showstopper”), major, and minor (often cosmetic). A critical problem prevents subjects from completing a task, such as not finding a power button or an enter key. A major problem substantially delays the subject but can be overcome. A minor problem has minimal impact on performance, but a change is nonetheless desired, such as making a label a different color or choosing a different font. For this study, task time was used to determine severity. Task times greater than 300 s (5 minutes) were critical. Greater than 30 s but less than or equal to 300 s were major. Minor was 30 s or less.

In the literature are a number of formal methods to analyze verbal protocols some of which the authors were unaware of at the time this experiment was conducted (e.g., [28]). However, for the purpose of this applied analysis, given the degree to which the subjects were expressive the experience of the experimenter, and the time available, a custom categorization scheme seemed more appropriate. In brief, problems were categorized as device domain or subject domain problems. Device domain problems were (1) visual or auditory interface related (layout, label, text, sound, and action time), (2) logic and organization (controls, search, system, and information architecture), or (3) nonusability software issues (stability, database, and response time). Subject problems included data knowledge, procedure knowledge, and preference. In this case, the scheme was to aid Korean designers, many of who were not fluent in English, and had no human factors background.

Finally, given focus on what the sponsor needed to fix, the limited time schedule, limited funding, no effort was made to examine the effect of subject differences (young versus old, those who had navigation systems versus those who did not, etc.) on the particular problems encountered. Although very interesting, they were secondary issues.

3.2. How Often Were Subjects Able to Complete the Tasks? Table 2 shows how well subjects did when they were given hints. One could interpret this data to suggest that the interface was relatively easy to use, but that was not the case. *Many subjects required hints to complete many of the tasks.* Had hints not been provided, the success rate would have been half the values shown, or typically less than 50%, quite poor. For perspective, subjects took 2 hours on average to complete these 20 tasks, which is about 6 minutes/task, a long time. Several of the tasks, such as setting a preset radio frequency, should only take a few seconds to complete.

3.3. How Often Did Problems Occur and How Severe Were They? Table 3 lists the 64 problems, a rather large number, from the most frequent to the least frequent. About two-thirds of all problems were experienced by at least 2 subjects. Among those, 16 problems (25 percent of the total) were encountered by more than half of all subjects. In the report summarizing this project [24] the frequency of problems was reported many ways. For the table that follows, they are reported at the number of subject (out of 13) experiencing a problem because this was the format most readily understood by the sponsor and its interface designers. Furthermore, the experiment was deliberately designed to minimize repetition

TABLE 2: Successful completion with hints.

#	Task	Success		# Not do
		#	%	
1	Set destination to 100 north 5th avenue, Ann Arbor, MI.	9	69	
2	Find the nearest McDonalds and set it as a waypoint.	8	62	
3	Find the nearest gas station from the current position.	11	85	
4	Cancel the route that you have set up.	10	77	
5	Enter and set a destination to the main Detroit Airport.	12	92	
6	Change map scale to zoom in slightly.	10	77	
7	Enter and set a destination to 1600 Pennsylvania Ave, Washington DC.	10	77	
8	Set the destination to the intersection of US-23 and I-94.	1	8	
9	Set the destination to Cobo Hall in Detroit, MI.	9	100	4
10	Set the destination to intersection of Broadway St and Plymouth Rd, Ann Arbor.	9	100	4
11	Enter your home address and store it.	3	50	7
12	Insert the CD and select the third track.	13	100	
13	Adjust volume to a comfortable level.	13	100	
14	Tune FM radio to 91.7.	13	100	
15	Adjust volume to a comfortable level.	13	100	
16	Save the frequency as first preset.	6	46	
17	Tune XM radio to Elvis station.	10	77	
18	Turn off XM radio.	12	100	1
19	Change the time to 11:45 pm.	13	100	
20	Connect the iPod and select the song "Wonderful Tonight."	12	92	

Note: under the heading success, # refers to the number of subjects out of 13 who successfully completed the task with hints. Thus, for task 1, $100 \times 9/13 = 69\%$. The # Not Do refers to the number of subjects who were not asked to do up to four tasks to keep the experiment from taking too long. Thus, for task 9, only 9 of the 13 subjects performed the task, but all 9 of them (100%) complete the task.

of tasks and task elements, and thereby more widely explore the interface, which tend to limit repeated encounter with problems.

The linkage between what was observed and these problems can best be described by example. Following is a description of some of the most frequent and critical problems. Problems were identified by a combination of how long it took to complete each task and step (if they were completed at all), what subjects did, what subjects said, and although not described here, their facial expressions. The ultimate indication of a problem was when they said, "I give up." Quite frankly, identifying when a problem had occurred was fairly obvious.

The most frequent problem (9) was that the system did not accept "DC" as a state name when subjects searched for an address in Washington, D.C (task 1). All subjects except one (92% of the sample) tried to type "DC." They would get to the state field and type in "D." Immediately, the C key would gray out because the system was expecting the subject to type "District." "Why is the C key gray? I want to type it." In contrast to most state names, the District of Columbia is invariably abbreviated as "DC," but the software accepted only "District of Columbia." Allowing the intelligent speller to accept both complete names and two-letter abbreviations would eliminate this problem, and speed the entry of other state names as well.

The system often froze (problem 10) while the experiment was in progress (for 11 of the 13 subjects, and for some subjects, multiple times). Given that a prototype was being evaluated, some problems were expected. The work around the freeze was to unplug the system, plug it back in again, and wait for the system to restart. This process took a minute or so to complete and was mildly annoying to subjects and experimenters. Although testing could have been conducted when the interface had fewer bugs that froze the interface, that testing would have occurred later in design when there was less time and fewer resources to correct problems identified.

For task 16, 10 of the 13 subjects could not figure out how to set the radio presets (push and hold the soft button), problem 42. When they reached the preset screen, they would say something such as "Where is the set button?" Often, they would push "autostore," which reset all the presets. In fact, one subject did this three times, exclaiming, "Why did it do that, again!" when the desired frequency setting appeared on a different button than it was prior to pressing autostore. Subjects did not realize that the method used to preset a radio frequency that works for radios with mechanical buttons (push and hold), also worked here.

Some 9 of the 13 subjects had trouble understanding the label to change the keyboard from alphabetic mode (the default) to numeric mode (problem 2), a step required to

TABLE 3: List of problems by frequency for all 13 subjects.

Frequency	Problem #	Severity	Description
12	9	Major	State name cannot be abbreviated (DC, Washington DC)
11	10	Minor	System freeze
10	14	Minor	No way to enter highway intersection as destination
10	42	Major	No intuitive method to save radio frequency as preset
9	2	Major	Difficult to switch keyboard to number keys
9	20	Major	Could not search for an XM channel by name, so they browsed the numeric listing, which was slow
9	24	Critical	Typed building number in the street name (north/no)
9	35	Major	Searching for an iPod song by title was very time-consuming
8	4	Minor	Guidance voice quality was inadequate
8	6	Critical	No intuitive method to cancel route/destination on map screen
8	7	Major	Buttons are easily confused with displays
8	8	Major	Expected this interface to support multitouch, in particular 2-finger zooming
8	17	Minor	Correct address rejected (wrong house number range)
7	28	Major	Did not realize what the search button did (also used for enter)
7	29	Major	Incomplete and insufficient text information
7	37	Critical	No alphabetic input method for numeric street name (Fifth Ave)
6	3	Minor	Negative response to autosuggest keyboard (autocomplete function)
6	13	Major	Did not realize there were multiple pages for destination screen
6	32	Major	Confusing intersection search guidance ("X", "space")
6	36	Major	No intuitive method to store the entered address/home address
6	45	Major	Did not understand icons on the map window (Z, compass, flag, current position, scale)
5	23	Major	Typed building number in the street name
5	26	Major	Did not understand the acronym POI
4	5	Major	Did not understand POI symbol/function on the map
4	16	Minor	Search displayed multiple entries for the same destination (3 Ann Arbor)
4	22	Major	Difficult to switch to alphabet keys
4	25	Critical	No intuitive method to search for a waypoint and add it as a waypoint on the map screen.
3	11	Minor	Too slow/fast system response time
3	12	Minor	Address/text information was not legible
3	27	Major	Did not recognize scale button on the map screen
3	30	Major	No way to back up (e.g., main mode screen, simulation mode, and local POI)
3	33	Minor	No intuitive method to bring up CD control after inserting CD
3	34	Major	Did not understand label SAT
3	38	Major	Did not understand label Zagat
3	39	Major	Did not recognize back button
3	40	Minor	Unable to get to CD mode after inserting CD
3	41	Major	Did not understand Autostore
3	44	Major	Did not understand icons on the status bar
3	57	Major	No intuitive graphic to select the searched result on the list (city/street)
2	46	Minor	Scale window disappears too quickly on the screen
2	48	Major	Narrow search coverage by POI Name (no support for McDonalds)
2	50	Major	Did not recognize the delete button on keyboard
2	64	Major	Did not understand label Map key
1	1	Minor	State name is highlighted and shown on Search by Address screen.
1	15	Minor	Confusing search result (Ann Arbor/Ann Arbor Twp)

TABLE 3: Continued.

Frequency	Problem #	Severity	Description
1	18	Minor	Misunderstood seek/track button label
1	19	Minor	Did not notice volume label
1	21	Major	No intuitive distinction between Route and Dest key
1	31	Minor	Wrong screen instruction ("search by POI" in intersection)
1	43	Minor	Did not understand label Dest
1	47	Minor	Select button location was inconsistent
1	49	Minor	Did not understand "Alternative" menu
1	51	Minor	Misunderstood Asian, Korean, and Chinese on POI list (as language)
1	52	Minor	No way to control volume on the screen
1	53	Minor	Did not understand preset scan
1	54	Minor	Misunderstood ATM and Bank on POI category list
1	55	Minor	Wrong unit information/temperature
1	56	Major	Did not understand play/stop icon
1	58	Major	Did not understand numbers next to the searched street names
1	59	Major	Not able to set clock in expected manner
1	60	Major	No intuitive layout for house number input
1	61	Minor	Did not understand orange line on the map
1	62	Minor	No autosave feature in entering city
1	63	Minor	Confusing categorization on POI list (gas station-travel/automotive)

Note: Problem #: The 64 problems were numbered 1–64.

complete task 1. They did not realize there was a key on the alphabetic keyboard screen to change modes, so they would search for other screens using the back key, proclaiming something like "where is the number screen?" At other times, they could say, "I do not know where it is, so I will try everything," and they selected each key on the screen one by one to learn what each did. Some subjects who were methodical in their efforts found the alphabetic keyboard mode key in this manner. The source of the problem was that the mode keys/buttons looked like other buttons on the keyboard, and there was no spacing or graphics to group them apart. There were other buttons for which functional grouping, indicated by spacing and graphics, would have helped as well.

In fact, subjects had numerous problems with soft buttons. First, there was no graphic distinction between buttons and displays, so that when subjects were not sure what to do, they pressed everything, including displays. Providing buttons with a drop shadow or other unique graphical characteristic as well as auditory feedback (a click sound) when buttons are operated should reduce confusion. When driving, drivers should not be looking at the display. Many interfaces have a beep to confirm that a switch has been pressed, but that beep is often the same as the beep for an error, which confuses drivers.

Also noteworthy were two problems related to street address entry. Many American street names have a direction as part of the name (e.g., North Main) and subjects were uncertain if they should enter the street name as North Main, Main and select North, or abbreviate North as N or NO. Only one option was provided, so only 9 of the 13 subjects were

able to complete entry of a street name containing a direction, even with hints. Often they tried using exactly the same and steps multiple times, saying something such as, "I think this is the way to do it, but I must have not done it quite right." They did not realize the system would not accept the data as they entered it.

Subjects also struggled with street names that contained numbers (e.g., 5th street) as only one method of entry was supported, even though subjects may choose to enter those streets as numbers or alphabetically (Fifth). Numbered streets are quite common in the United States.

Related to this were problems associated with subjects not realizing what had been set or was a default. This was particularly true with setting the state. Keep in mind that many cities are located on rivers, because rivers provide both water and transportation. However, rivers also serve as a geographical boundary, so going to a nearby place may require changing the state.

Problems in search for songs and radio stations were common. In part, this was because the interface for each mode (AM/FM, XM, etc.) was unique. The criteria on which one could search was unique to the mode, and most importantly, so was the organization. What could be saved or preset varied with the mode, including what saves or presets were named. As a consequence, subjects needed to browse through pages of screens to find an XM radio station or song on a storage device, which was a very time-consuming process. Manually scrolling through a list of 100 or more items and reading them to find a desired item should not be done while driving, especially when the lists are in an order the subject cannot use to speed the search.

In summary, from most to least, problems included unclear labels (20 instances, such as that for the number key and the name autostore), problems with search (10, such as a lack of consistency in method names and methods available, especially in finding songs and XM stations), poor graphics (9, many icons were meaningless), disorganized system (9, information such as destination modes being split across 2 screens), illegible text (5, mostly text that was too small, especially on maps), poor layout (4, such as inconsistent location of the “ok” and “done” buttons), other organizational issues, unreliable software, and database errors (2 each, including missing addresses), and problems associated with unrecognizable sounds, slow system response (2 types), and disorganized controls (all 1 each).

3.4. Persistence. Using persistence to identify problems was less useful here than has been identified in the literature. Examples of the most persistent problems included not understanding the acronym POI, not understanding icons in the map window, system freeze, not knowing what the search button did, expecting multitouch to be supported, and not understanding the label Zagat. If anything, the persistence data reinforced the need for better icons and graphics, and potentially eliminating icons in some cases. As an example, there were two screens from which subjects could select a method to enter a destination. Often, they did not realize there was a second screen, so they were stuck. The icons were of no help. Had the icons been removed, leaving only text, all of the entry methods would be on one screen and user performance would likely have improved.

Although there are numerous navigation systems in use today, and many of them use icons, there are no standard icons for navigation functions in ISO Standard 2575 [29]. Although it may not be possible to develop well-understood icons for many of the navigation functions of interest, whatever is developed could be better than the current situation, where icons vary from system to system.

3.5. Combined Analyses. When presented with a huge list of problems, such as those in Table 3, the designers’ immediate reaction is often to ignore the overwhelming user feedback. First, there is disbelief that subjects experienced all of the problems listed. Therefore, two representatives from the sponsor responsible for the interface design observed every subject, so they saw that the problems were real. Unfortunately, they were not native English speakers, so for the first few subjects they struggled to understand what was occurring. That was overcome by impromptu discussions with them between test sessions or at the end of the day to explain what was observed. Also provided were a few video outtakes for others not present. In a subsequent project, a secure web camera in the test room was provided so those not present could observe the experiment. In this case, the 13-hour time difference between the test facility (Ann Arbor, Michigan, USA) and the sponsor’s main engineering center (Yongin-Shi, Gyunngi-Do, Korea) makes remote viewing inconvenient.

TABLE 4: Problem numbers listed summarized by frequency and severity.

Frequency	Severity		
	Critical ($t > 300$ s)	Major ($300 \geq t > 30$ s)	Minor ($30 \geq t$)
12		9	
11			10
10		42	14
9	24	2, 20, 35	
8	6	7, 8	4, 17
7	37	28, 29	
6		13, 32, 36, 45	3
5		23, 26	
4	25	5, 22	16
3		27, 30, 34, 38, 39, 41, 44, 57	11, 12, 33, 40
2		48, 50, 64	46
1		21, 56, 58, 59, 60	1, 15, 18, 19, 31, 43, 47, 49, 51, 52, 53, 54, 55, 61, 62, 63

Note: for the descriptions corresponding to the problem numbers listed here, see Table 3.

To help designers prioritize what they should do, tables were created using pairs of the dimensions of interest (frequency, severity, and persistence). Ideally, designers should consider those dimensions, the effort required to fix each problem, and the implications of fixing each problem on other problems, as well as other factors in making decisions about what to fix.

The frequency-severity table (Table 4) contains the most useful of the dimension combinations. The problems listed in the upper left area of the table (e.g., 9, 42, 24, 2, 20, and 35) are the most frequent and severe, and therefore, of the highest priority.

4. Conclusions and Discussion

There were many problems with this interface, some of which were expected early in design when this interface was examined. Subjects consistently needed hints to complete tasks, which was not indicative of an intuitive interface.

One might criticize the sponsor of the research for the existence of these problems. However, the more important point is that they supported qualified experts to examine their interface, identify problems, and suggest improvements. Keep in mind that it is not the researchers’ role to make the changes desired, and what is changed represents a tradeoff between user impact, cost, schedule, and hardware and software limitations.

4.1. Lack of Style Guide. There were inconsistencies in the interface, for example, where the “done” or “ok” button was located (and how it was labeled). According to the sponsor’s representatives, there was no style guide or other specific set

of guidelines governing the interface, a situation that greatly increases the likelihood of problems due to inconsistency in the user interface. Creating a style guide, especially one based on research, is a major task, but well worth the effort. Most computer manufacturers have style guides to help ensure their interfaces are consistent (e.g., [30]).

4.2. Inconsistent Music Interfaces. The search methods available and what could be stored as presets varied with the media and are reflected in problems 16, 20, 28, 35, and others. These inconsistencies led to interfaces that were unique to each media, making interface navigation difficult. Admittedly, the underlying databases have different structures, but a more common format and more similar search features would have been beneficial, so that subjects would only need to learn one set of search methods that were consistently named, not a unique set of methods for each data set. Interestingly, this system (and most others) did not allow for aggregation of all favored music presets (AM/FM, XM, etc.) on a single screen.

4.3. Destination Entry Problems. There were numerous issues with destination entry, most due to limiting the ways in which information was to be entered and not making apparent what had already been set. These issues are reflected in problems 14, 16, 17, 37, and others.

4.4. Icons Were Not Understood. There was a desire to provide icons so the interface would be language independent and usable by a wider user group. That assumes the icons are understandable, which was not true here. Although better icons could be developed and included in ISO Standard 2575, how well they will be understood is uncertain. That suggests for some parts of the interface, icons may not be provided. In the case of selecting a destination entry method, all of the methods will fit on one screen instead of being distributed across two screens, making it easier for users to find the desired method. In other cases (e.g., maps), icons must be provided, as there is insufficient space for text (let alone icons plus text). However, in this instance, most icons were not understood (problems 44, 45, 64, etc.). In fact, in a very lengthy examination of the icons used in this interface, conducted after all tasks were completed, on average only 2 of the 13 subjects were able to correctly identify what the various icons meant when shown in context. Easy to understand map icons need to be developed.

4.5. Labels Were Not Understood. This included POI (problem 5), SAT (problem 34), Zagat (problem 38), and others. Each of the labels used should be considered and alternatives proposed. This needs to be done in conjunction with the effort to develop new icons as they are an alternative to text labels.

4.6. Lack of Functional Grouping. There were several instances where information on screens was not grouped by function, increasing the time for users to find particular information (and increasing errors as well). The best example of this is problem 2. Admittedly, space is extremely limited,

but there were instances where spacing and graphics could have been utilized for this purpose.

4.7. Buttons and Displays Looked Alike. There were no common graphical elements to controls (primarily soft buttons) and displays (mostly icons), so that when subjects were lost, they pressed everything (problem 7). In such instances, having different auditory feedback for operation of a button and erroneous operation would have been helpful.

Many of these problems are not new (See [31]).

Beyond this specific interface, which does this experiment say about how think-aloud experiments should be conducted?

4.8. More Experimenter Training Needed. More time was needed to train the experimenter in the think-aloud method, in particular time spent on testing pilot subjects and reviewing video recordings of them. In this project, there were two days between when the interface actually worked and when testing had to begin to deliver results to meet the sponsor's production schedule, far too short. A minimum of two to three weeks is recommended. Training is particularly important for nonnative speakers of the language of subjects or those who do not have extensive experience in testing human subjects as experimenters. Typically, they do not prompt subjects enough, or more generally, just have problems in reading subjects, not knowing when to engage them. This need was reflected in silent periods, where neither the subject or the experimenter spoke, and often the experimenter just stared at the screen. In other situations, when designers without human factors/usability expertise conduct the testing, excessive leading of subjects is observed, typically involving telling subjects how to complete a task.

A list of probe questions and criteria for when those questions should be asked would be a useful addition to the training materials. Being repeatedly asked "what are you thinking" is annoying. Formal rules about when to intervene and what to do (just press this button, but not saying why) can be helpful.

Another idea is a device that would look for periods of silence and subject inaction. When those periods occurred for some time, the device would vibrate something on which the experimenter was sitting as a reminder to ask a probe question.

4.9. Real-Time Session Recording Helped. Secretaries or students who were fast typists sat in the room with the subject and the experimenter and served as the scribe, creating the session transcripts. Immediately after the session, when the session was fresh in their mind, the scribe checked their transcript against the audio and video recording. For this to occur, more than a few minutes is required between subjects. Nonetheless, generating transcripts in real time rather than after the fact from a video recording reduced the time to provide the results to the sponsor, which was important where a real evolving product with a rigid development schedule was concerned.

There was no evidence that the presence of a scribe was disruptive, which was a concern. For example, subjects

did not look at the scribe or comment on what they were doing. Being nearby, the scribe could hear when the subject mumbled, which was often the case when people think aloud. The mumbling is often difficult to hear on recorded video, but is often the most informative part of an interaction. In this experiment, no software to support recording was used. Use of the Morae software to aid session recording [32] should be considered for the future.

There was also no time to acquire and set up a high-quality audio recording system. Thus, due to substandard audio quality, few segments could be used for outtakes. To partially overcome this problem, English subtitles of what the subjects and experimenter said were added. Time and cost permitting, even better for this audience would have been subtitles in both English and Korean.

4.10. Real-Time Observation Helped. The project team has always invited sponsors to watch test sessions directly. There are always concerns that additional observers will distract subjects, or they might do or say something that will interfere with a test protocol, requiring the data to be discarded (and additional time for replacement data to be collected). In on-road studies, seating for the subject, the experimenter, and equipment may leave only one unoccupied seat.

In this instance, there were two observers from the sponsor who sat quietly in the corner of the test room. Each night, they produced a summary that they sent to their designers, greatly shortening the time to provide feedback to them. The authors could not have responded as quickly. This feedback was an extremely important supplement to the written report and videos.

Had the hardware been available, providing a real time, web camera video of the experiment to the designers (in Korea) would have been useful. If the website URL is not advertised and a website is password protected, security should not be an issue. In this case, there were also issues with test sessions being a half-day out of synch with that of the designers and uncertainties about their ability to understand spoken English when it is disjointed and mumbled.

Surprisingly, the major challenge in getting the video to the users was not associated with the source, but getting the feed into the sponsor network because of security constraints. The solution is for the sponsor's employees to stay home when testing is expected, and watch it from there.

Keep in mind that if remote viewing is implemented, permission must be obtained from subjects on the consent form (and from the human subject board when the experiment is reviewed).

4.11. Data Reduction Was Slow. Creating a list of problems was a very slow and labor-intensive process, requiring the second author to read the transcripts and watch each session many, many times. Had there been time, the classification of problems would have benefited from a more structured approach, either considering them as problems related to goals and methods or to various error types (e.g., [33, 34]). A more structured approach could have shortened the time for data reduction, but only by a small amount. (See [16, 35–37].) If anything, a major weakness of the think-aloud method

is that it takes so long to reduce the data, a concern when product development needs to be rapid.

4.12. More Information Identifying Solutions Was Needed. As the focus of the project was on identifying problems, less effort was given to identifying solutions. Had the time and resources been added, having a table that identified the solution to each problem (if there was a solution) could be useful.

The tone of these last comments may suggest the think-aloud method is seriously flawed. Quite to the contrary, the method provided extensive and detailed insights as to why users struggled to use the prototype navigation radio and how it could be improved. Wider use for automotive interface evaluation is strongly encouraged. Much of the information provided in this paper either could not have been obtained or would be difficult to obtain using other methods.

The ultimate test of this experiment was how the results were used. After being presented to Mobis, the sponsor, there was a follow-up presentation for Hyundai-Kia. The result was numerous requests from Hyundai-Kia to Mobis to modify the prototype interface to enhance its usability based on the experimental results. This was not a project that led to a report that just sat on a shelf. Furthermore, Hyundai-Kia has funded follow-on research now in progress.

Conflict of Interests

The authors declare that they have no conflict of interests.

Acknowledgment

This research was supported by a contract from Mobis to the University of Michigan.

References

- [1] J. Nielsen and R. Mack, *Usability Inspection Methods*, John Wiley & Sons, New York, NY, USA, 1994.
- [2] S. Krug, *Rocket Surgery Made Easy: The Do-It-Yourself Guide to Finding and Fixing Usability Problems*, New Riders, Berkeley, Calif, USA, 2010.
- [3] J. Rubin, D. Chisnell, and J. Spool, *Handbook of Usability Testing*, John Wiley & Sons, New York, NY, USA, 2008.
- [4] R. Jeffries, J. R. Miller, C. Wharton, and K. Uyeda, "User interface evaluation in the real world: a comparison of four techniques," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 119–124, 1991.
- [5] R. A. Virzi, J. F. Sorce, and L. B. Herbert, "Comparison of three usability evaluation methods: Heuristic, think-aloud, and performance testing," in *Proceedings of the 37th Annual Meeting the Human Factors and Ergonomics Society*, pp. 309–313, October 1993.
- [6] R. Molich, M. R. Ede, K. Kaasgaard, and B. Karyukin, "Comparative usability evaluation," *Behaviour and Information Technology*, vol. 23, no. 1, pp. 65–74, 2004.
- [7] K. Hornbæk, "Current practice in measuring usability: challenges to usability studies and research," *International Journal of Human Computer Studies*, vol. 64, no. 2, pp. 79–102, 2006.

- [8] M. W. M. Jaspers, "A comparison of usability methods for testing interactive health technologies: Methodological aspects and empirical evidence," *International Journal of Medical Informatics*, vol. 78, no. 5, pp. 340–353, 2009.
- [9] J. Nielsen and R. Molich, "Heuristic evaluation of user interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '90)*, Association for Computing Machinery, New York, NY, USA, 1990.
- [10] A. M. Lund, "Expert ratings of usability maxims," *Ergonomics in Design*, vol. 5, no. 7, pp. 15–20, 1997.
- [11] E. Hvannberg, E. Law, and M. Lárusdóttir M, "Heuristic evaluation: comparing ways of finding and reporting usability problems," *Interacting with Computers*, vol. 19, no. 2, pp. 225–240, 2007.
- [12] C. H. Lewis and Using the, "'Thinking Aloud" method in cognitive interface design," Technical Report IBM RC-9265, IBM Watson Research Center, Yorktown Heights, NY, USA, 1982.
- [13] P. C. Wright and A. F. Monk, "The use of the Think-Aloud evaluation methods in design," *ACM SIGCHI Bulletin*, vol. 23, no. 1, pp. 55–57, 1991.
- [14] E. Krahmer and N. Ummelen, "Thinking about thinking aloud: a comparison of two verbal protocols for usability testing," *IEEE Transactions on Professional Communication*, vol. 47, no. 2, pp. 105–117, 2004.
- [15] M. Norgaard and K. Hornbaek, "What do usability evaluators do in practice? An exploratory study of Think-Aloud testing," in *Proceedings of the 6th conference on Designing Interactive systems (DIS '06)*, pp. 209–218, 2006.
- [16] J. Lewis, "Usability testing," in *Handbook of Human Factors and Ergonomics*, G. Salvendy, Ed., chapter 46, pp. 1267–1312, John Wiley & Sons, New York, NY, USA, 4th edition, 2012.
- [17] D. Manes, P. Green, and D. Hunter, "Prediction of destination entry and retrieval times using keystroke-level models," Technical Report UMTRI-96-37, University of Michigan Transportation Research Institute, Ann Arbor, Mich, USA, 1997, EECS-ITS LAB FT97-077.
- [18] P. Green, "Estimating compliance with the 15-Second Rule for driver-interface usability and safety," in *Proceedings of the 43rd Annual Meeting on Human Factors and Ergonomics Society*, Human Factors and Ergonomics Society, CD-ROM, Santa Monica, Calif, USA, 1999.
- [19] C. Nowakowski and P. Green, "Prediction of menus selection times parked and while driving using the SAE J2365 method," Technical Report 2000-49, University of Michigan Transportation Research Institute, Ann Arbor, Mich, USA, 2001.
- [20] Calculation of the Time to Complete in-Vehicle Navigation and Route Guidance Tasks, SAE Recommended Practice J2365, 2002.
- [21] S. Schneegass, B. Pfleging, D. Kern, and A. Schmidt, "Support for modeling interaction with automotive user interfaces," in *Proceedings of the the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '11)*, pp. 71–78, 2011.
- [22] B. T.-W. Lin, P. Green, T.-P. Kang, A. Mize, A. Best, and K. Su, "Touch screen menu selection time and SAE J2365 predictions of them," Technical Report UMTRI-2012-9, University of Michigan Transportation Research Institute, Ann Arbor, Mich, USA, 2012.
- [23] J. Sauro, "Margins of error in usability tests," 2009, <http://www.measuringusability.com/test-margin.php>.
- [24] J.-S. Park, P. Green, and M. Alter, "Usability assessment of the mobis gen 3 navigation radio," Technical Report UMTRI-2010-23, University of Michigan Transportation Research Institute, Ann Arbor, Mich, USA, 2010.
- [25] J. R. Lewis, "Sample sizes for usability tests: mostly math, not magic," *Interactions*, vol. 13, no. 6, pp. 29–33, 2006.
- [26] R. A. Virzi, "Refining the test phase of usability evaluation: how many subjects is enough?" *Human Factors*, vol. 34, no. 4, pp. 457–468, 1992.
- [27] J. Sauro and J. R. Lewis, *Quantifying the User Experience*, Morgan Kaufmann, San Francisco, Calif, USA, 2012.
- [28] Y. Kinoe, "The VPA method: a method for formal verbal protocol analysis," in *Designing and Using Human-Computer Interfaces and Knowledge Based Systems*, G. Salvendy and M. J. Smith, Eds., pp. 735–742, Elsevier Science, Amsterdam, The Netherlands.
- [29] Road Vehicles—Symbols for Controls, Indicators, and Tell-tales, ISO Standard 2575, 2010.
- [30] Apple Inc., "OS X Human Interface Guidelines", Apple, Cupertino, Calif, USA, 2012, <http://developer.apple.com/library/mac/documentation/userexperience/conceptual/appleguidelines/OSXHIGuidelines.pdf>.
- [31] C. Nowakowski, P. Green, and O. Tsimhoni, "Common automotive navigation system usability problems and a standard test protocol to identify them," in *Proceedings of the the ITS-America Annual Meeting (CD-ROM)*, Intelligent Transportation Society of America, Washington, DC, USA, 2003.
- [32] Techsmith, "Morae (software)," Techsmith, Okemos, Mich, USA, 2012, <http://www.techsmith.com/morae.html>.
- [33] J. Reason, *Human Error*, Cambridge University Press, Cambridge, UK, 1990.
- [34] S. A. Shappelland and D. A. Wiegmann, "The human factors analysis and classification system-HFACS," Final Report DOT/FAA/AM-00/7, Office of Aviation Medicine, Federal Aviation Administration, US Department of Transportation, Washington, DC, USA, 2000.
- [35] T. S. Andre, H. Rex Hartson, S. M. Belz, and F. A. McCreary, "User action framework: a reliable foundation for usability engineering support tools," *International Journal of Human Computer Studies*, vol. 54, no. 1, pp. 107–136, 2001.
- [36] K. Hornbaek and E. Forkjaer, "Comparison of techniques for matching usability problem description," *Interacting With Computers*, vol. 20, no. 6, pp. 505–514, 2008.
- [37] R. Khajouei, L. W. P. Peute, A. Hasman, and M. W. M. Jaspers, "Classification and prioritization of usability problems using an augmented classification scheme," *Journal of Biomedical Informatics*, vol. 44, no. 6, pp. 948–957, 2011.