



Decision Models for Comparative Usability Evaluation of Mobile Phones Using the Mobile Phone Usability Questionnaire (MPUQ)

Young Sam Ryu

Ingram School of Engineering
Texas State University-San
Marcos
San Marcos, TX 78666, USA

Kari Babski-Reeves

Department of Industrial &
Systems Engineering
Mississippi State University
Starkville, MS 39762, USA

Tonya L. Smith-Jackson

Grado Department of
Industrial & Systems
Engineering
Virginia Tech
Blacksburg, VA 24061, USA

Maury A. Nussbaum

Grado Department of
Industrial & Systems
Engineering
Virginia Tech
Blacksburg, VA 24061, USA

Abstract

A comparative usability evaluation was performed using various subjective evaluation methods, including Mobile Phone Usability Questionnaire (MPUQ). Further, decision-making models using Analytic Hierarchy Process (AHP) and multiple linear regression were developed and applied. Although the mean rankings of the four phones were not significantly different across the evaluation methods, there were variations across the methods in terms of the number of rank orderings, preference proportions, and methods to select their initial preference. Thus, this study provided a useful insight into how users make different decisions through different evaluation methods. Also, the result showed that answering a usability questionnaire affected a user's decision-making process for comparative evaluation.

Keywords

usability, mobile phone, questionnaire, multi-criteria decision-making method, linear regression, Analytic Hierarchy Process, Mobile Phone Usability Questionnaire, Post Study System Usability Questionnaire

Introduction

The MPUQ (Ryu & Smith-Jackson, 2006) can be used to evaluate usability of mobile phones for the purpose of making decisions among competing phone variations in the end-user market, developing prototype alternatives during the development process, and evolving versions of a phone during an iterative design process. The typical decision-making method using the questionnaire requires averaging the score of all 72 questions. Although Nunnally (1978) pointed out that the effort in developing weights typically does not have much of an effect on a scale's reliability, validity, or sensitivity, the method of simply averaging the item scores is unlikely to reflect the way people make decisions. Therefore, alternative methods for determining usability scores may provide useful insight into the decision-making process.



The manner in which humans make decisions varies considerably across individuals and situations. Early research on decision-making theory focused on the way humans were observed to make decisions (descriptive) and the way humans should theoretically make decisions (normative). Although the distinction between descriptive and normative models has become fuzzy, it is important to clearly distinguish between them because the distinction can be a useful reference point in attempting to improve decision-making processes (Dillon, 1998). In addition, prescriptive models have been introduced and are based on the theoretical foundation of normative theory in combination with observations of descriptive theory. However, some researchers use “normative” and “prescriptive” interchangeably (Bell, Raiffa, & Tversky, 1988b). As a way of distinguishing the three different models of decision-making, Table 1 shows the taxonomy for classification.

Table 1. Classification of decision-making models (Bell, Raiffa, & Tversky, 1988a; Dillon, 1998)

Classifier	Definitions
Descriptive	What people actually do, or have done
	Decisions people make How people decide
Normative	What people should and can do
	Logically consistent decision procedures How people should decide
Prescriptive	What people should do in theory
	How to help people to make good decisions How to train people to make better decisions

The most prominent distinction among different decision-making theories or models is the extent to which they make trade-offs among attributes (Payne, Bettman, & Johnson, 1993); classifying models as either non-compensatory or compensatory. A non-compensatory model is any model in which “surpluses on subsequent dimensions cannot compensate for deficiencies uncovered at an early stage of the evaluation process; since the alternative will have already been eliminated” (Schoemaker, 1980, p. 41), while a compensatory model occurs when “a decision maker will trade-off between a high value on one dimension of an alternative and a low value on another dimension” (Payne, 1976, p. 367). Among the three different decision-making models (Table 1), the descriptive models are considered non-compensatory, while the normative and prescriptive models are typically regarded as compensatory (Dillon, 1998).

The goal of this research was to provide greater sensitivity of the response of MPUQ for the purpose of comparative usability evaluation and to determine which usability dimensions and questionnaire items contribute most to making decision regarding the preference of mobile phones. In a previous paper, the Analytic Hierarchy Process (AHP) was used to develop normative decision models to provide composite scores from the responses of MPUQ (Ryu, Babski-Reeves, Smith-Jackson, & Nussbaum, 2007). In this paper, multiple linear regression was employed to develop models to provide composite scores from the responses of MPUQ as well.

Study 1: Development of Regression Models

Method

Four different models of mobile phones were evaluated for overall usability. A within-subject design was used to study the effect of phone model (4 levels) on usability ratings to reduce the variance across participants.

Equipment

Four mobile phone models from different manufacturers were provided as the evaluation targets, each having the same level of functionality and price range (\$200-\$300), along with user’s manuals. Identification letters were assigned to each phone (A to D) and the letter

stickers were placed on the brand names on the phone to minimize the exposure of the brand names of the phones.

Participants

The 16 participants, eight Minimalists¹ and eight Voice/Text Fanatics², were recruited. None of the participants owned or had owned one of the four phones.

Procedure

A participant was assigned to a laboratory room provided with the four different mobile phones. The participant was asked to complete a predetermined set of tasks on each phone. The tasks were those frequently used in mobile phone usability studies:

1. Add a phone number to phone book.
2. Identify the last outgoing call stored in the phone, including name and phone number.
3. Set an alarm clock to 7 AM.
4. Change current ringing signal to vibration mode.
5. Change the current ringing signal from vibration mode to the sound you like.
6. Send a short text message.
7. Send a text message 'Hello World!' to ###-###-####.
8. Take a picture of this document and store it.
9. Delete the picture you just took.

This session was intended to provide a basic usage experience with each phone to provide a basis by which to answer the questionnaire, and to standardized usage knowledge for each phone.

After completing this session, participants provided a score from 1 to 7 to determine the ranking of each phone regarding the preference (post-training [PT]). Thus, the score was used as the dependent variable in the regression model.

For the evaluation session using the MPUQ, participants completed all the questionnaire items for each phone according to a predetermined counterbalanced order. Each participant was allowed to explore the phones and perform any task. There was no time limit to complete the session.

Independent variables were to be responses on a Likert-type scale from 1 to 7 for each question of the MPUQ. Since each participant provided an absolute score on each phone, there were four observation points per participant. Thus, there were 32 observations for each user group of Minimalists and Voice/Text Fanatics. The MPUQ consisted of 72 questions, so that the number of observations was not enough to generate regression models if all the 72 questions were used as independent variables separately; the observation number should be at least larger than the number of independent variables. One reasonable way to deal with this limitation was to use each factor as one independent variable. The 72 questions were grouped into six different categories by the factor analysis in Ryu and Smith-Jackson (2006):

- Ease of learning and use (ELU)
- Assistance with operation and problem solving (AOPS)
- Emotional aspect and multimedia capabilities (EAMC)
- Commands and minimal memory load (CMML)
- Efficiency and control (EC)
- Typical tasks for mobile phones (TTMP).

Thus, 32 observations were reasonably sufficient to develop a regression model having six independent variables. Response data from the 72 questions of the MPUQ were combined into

¹ Users who employ just the basics for their mobility needs

² Users who tend to be focused on text-based data and messaging; Please see Ryu and Smith-Jackson (2006) for the description of mobile user group categorization by IDC (2003).

six factors, which were obtained by taking the arithmetic mean of the response on the questions of each factor.

Results

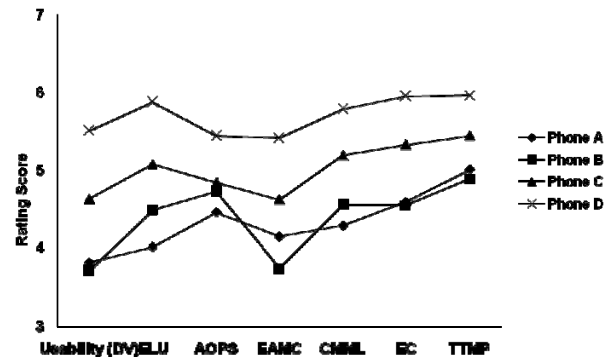


Figure 1. Mean scores of the dependent variable and independent variables for Minimalists

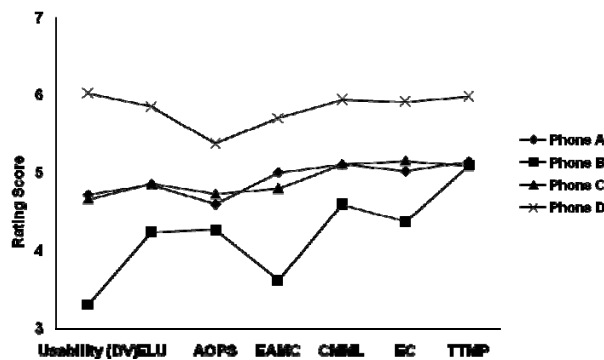


Figure 2. Mean scores of the dependent variable and independent variables for Voice/Text Fanatics

According to the descriptive statistics (Figure 1 and Figure 2), phone D was the most preferred phone for both user groups. Also, phone B showed the largest variation of scores among groups of variables for both user groups.

Table 2 and Table 3 show regression model statistics for the user groups. Both models showed good adequacy as evidenced by the adjusted R-square values and the models were highly significant (p-values less than 0.0001). However, model adequacy was higher for Voice/Text Fanatics (Adj R-Sq = 0.8632) than Minimalists (Adj R-Sq = 0.6800).

Table 2. Analysis of variance result of the regression model for Minimalists

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	61.09929	10.18322	11.98	<.0001
Error	25	21.24946	0.84998		
Corrected Total	31	82.34875			
Root MSE		0.92194	R-Square	0.742	
Dependent Mean		4.41875	Adj R-Sq	0.680	
Coeff Var		20.86433			

Table 3. Analysis of variance result of the regression model for Voice/Text Fanatics

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	73.78272	12.29712	33.61	<.0001
Error	25	9.14603	0.36584		
Corrected Total	31	82.92875			
Root MSE		0.60485	R-Square	0.8897	
Dependent Mean		4.68125	Adj R-Sq	0.8632	
Coeff Var		12.92065			

As the result of multiple linear regression analysis, each model provided an intercept value and six coefficients for six groups of variables (Table 4 and Table 5).

Table 4. Parameter estimates of the regression model for Minimalists

Variable	DF	Parameter Estimate	Standard Error	T Value	Pr > t
Intercept	1	-0.60783	1.33369	-0.46	0.6525
ELU	1	-0.00546	0.51098	-0.01	0.9916
AOPS	1	-0.43095	0.47680	-0.90	0.3747
EAMC	1	0.77836	0.26436	2.94	0.0069
CMMML	1	-0.38602	0.46432	-0.83	0.4136
EC	1	0.79477	0.57989	1.37	0.1827
TTMP	1	0.28423	0.25742	1.10	0.2800

Table 5. Parameter estimates of the regression model for Voice/Text Fanatics

Variable	DF	Parameter Estimate	Standard Error	T Value	Pr > t
Intercept	1	-1.0467	0.84670	-1.24	0.2279
ELU	1	1.32712	0.36306	3.66	0.0012
AOPS	1	0.81703	0.29001	2.82	0.0093
EAMC	1	0.09528	0.18705	0.51	0.6150
CMML	1	-0.55108	0.31206	-1.77	0.0896
EC	1	0.48106	0.36106	1.33	0.1948
TTMP	1	-0.89725	0.25147	-3.57	0.0015

EAMC was the only significant factor relating to phone selection for Minimalists ($p < 0.0069$), while ELU ($p < 0.0012$), AOPS ($p < 0.0093$), and TTMP ($p < 0.0015$) were significant for Voice/Text Fanatics. This is a very interesting result, since it provided insight into the most influential usability dimensions for each user group. The list of parameter estimates was used as the coefficients to produce composite scores (denoted as REG method) in Study 2.

Study 2: Comparative Evaluation with the Models

To validate the application of the Analytic Hierarchy Process (AHP)-applied MPUQ (Ryu, Babski-Reeves, Smith-Jackson, & Nussbaum, 2007) as well as the regression models of Study 1, a comparative usability evaluation of the four different mobile phones was conducted. Also, sensitivity analysis was performed comparing the AHP-applied MPUQ model, the MPUQ without the AHP model, and the decision model derived by linear regression. The population of participants was concentrated on the two identified majority groups (i.e., Minimalists and Voice/Text Fanatics).

Method

A within-subject design was used. This choice of a within-subject design is also compatible with the idea that users or consumers explore candidate phones to make decisions. Thus, each participant was given all the phones to evaluate. A completely balanced design was used to determine the order of exposure to each phone for evaluation. Therefore, each participant completed four sets of the MPUQ (one for each phone). Also, they completed four sets of the Post-Study System Usability Questionnaire (PSSUQ) (Lewis, 1995).

Equipment

The same four phones evaluated in Study 1 were provided. An identification letter was given to each phone, from A to D, to be referred to during the experimentation.

PSSUQ was selected as the existing usability questionnaire for several reasons. First, PSSUQ employs Likert-type scales with seven steps, which are the same specifications of the MPUQ. Hence, it is easy to compare the individual score of items and the overall score by averaging the item scores. For this reason, another candidate, Software Usability Measurement Inventory (SUMI), was excluded because it uses dichotomous scales. Second, PSSUQ has a relatively small number of items, 19, so it takes less time to complete than other questionnaires, such as SUMI (50 items) (Kirakowski & Corbett, 1993), Purdue Usability Testing Questionnaire (PUTQ) (100 items) (Lin, Choong, & Salvendy, 1997), and Questionnaire for User Interaction Satisfaction (QUIS) (127 items) (Harper & Norman, 1993).

Participants

Since there were four different mobile phones to be evaluated with a completely counter-balanced design, the number of participants was 24 (4!). A total of 48 participants was

recruited, because two user groups were to be evaluated. Also, all of them were non-users of the four phones evaluated.

Procedure

Participants were provided with the four phones, user guides, and the four identical sets of MPUQ and PSSUQ individually. Before the participant started the evaluation session, he or she was asked to rank his or her preferences for all the phones based on his or her first impression (FI). He or she was allowed to examine the phones briefly (for two minutes or less), then asked to determine the ranking of each phone regarding the preference.

Familiarity with each phone was provided by having participants complete a predetermined set of tasks on each phone (identical to that of Study 1). Participants then ranked each phone (post-training [PT]) on a 1 to 7 scale.

Participants completed the questionnaires (MPUQ and PSSUQ) for each phone following the familiarization session level. There was no time limit to complete the session. The order of the questionnaires was counter-balanced.

After answering both the MPUQ and PSSUQ, each participant was asked to rank order the phones again (post-questionnaires [PQ]) to determine if rankings changed after the participants completed the usability questionnaires. In other words, the usability evaluation activity required by the questionnaire may have affected the post-training decision.

Results

Mean Rankings

Seven different sets of ordered rankings on the four mobile phones were collected. The data sets were those from (1) first-impression ranking (FI), (2) post-training ranking (PT), (3) post-questionnaire ranking (PQ), (4) ranking from the mean score of the MPUQ (MQ), (5) ranking from the mean score of the PSSUQ (PSSUQ), (6) ranking from the AHP model of MPUQ (AHP), and (7) ranking from the regression model of MPUQ (REG). Thus, the treatments are the different phones (4 phones) and an observation consists of a respondent's ranking of the phones from the most preferred to the least preferred.

Table 6 shows an example of the data collected based on the FI for the Minimalists group in a ranked format. Since there were seven different methods to obtain the rankings for the two user groups, 14 tables similar to Table 6 were gathered. Based on the ranked data, the mean rank for each phone was obtained and charted (Figure 3 and Figure 4). In general, it was observed that the mean ranks of phones were phone D, phone C, phone A, and phone B in ascending order for both user groups, which is from the most favorable to the least favorable in interpretation. However, it seemed difficult to confirm that phone D received a greatly better rank from the Minimalists because the rank differences were so close by the FI, PT, PQ, and MQ. Also, it was observed that phones A and B received almost the same mean rank from the Voice/Text Fanatics group with the regression model. To investigate whether the mean rankings of the phones are significantly different, the Friedman tests were performed.

Table 6. Ranked data format by first impression

Participant	Rankings for Phones			
	A	B	C	D
1	3	4	2	1
2	4	3	2	1
.
24	4	2	3	1

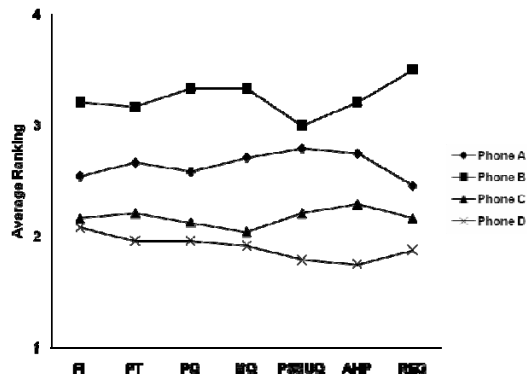


Figure 3. Mean rankings for Minimalists

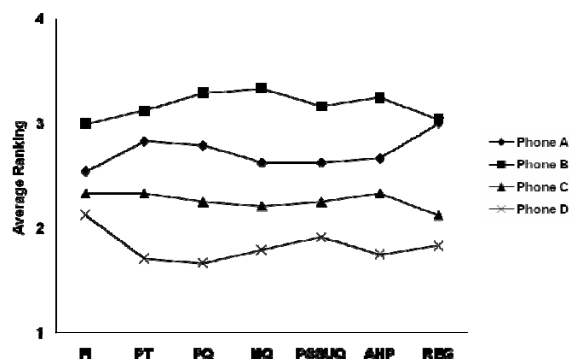


Figure 4. Mean rankings for Voice/Text Fanatics

Preference Data Format

The ranked data format could be converted to the preference data format suggested by Taplin (1997) to observe more information that is difficult to investigate with the mean rank of each phone. ABCD was used to denote the response where the participant most preferred A, next preferred B, next preferred C, and finally D. When multiple responses display the same ordering, this was represented by a number preceding the notation. For example, 3ABCD indicates that three participants ordered them in ABCD.

The summary of the preference data from all the seven methods of evaluation by each user group was presented (Table 7 and Table 8). PSSUQ provided the greatest number of different orderings, while AHP and MPUQ provided the smallest number of different orderings (Table 7). The ties are indicated by underscore in the preference format.

Table 7. Summary of the preference data from each evaluation method (Minimalists)

FI	PT	PQ	MQ	PSSUQ	AHP	REG
5DCAB	5DCBA	4DCAB	6DCAB	2BACD	7DCAB	6DCAB
3ACDB	3ACDB	3ACDB	5DCBA	2CCBA	3ACDB	4CDAB
3ADCB	3DCAB	3DCBA	3ACDB	2DACB	3DBCA	3ACDB
3CDBA	2ADCB	2ADCB	2CADB	2DBCA	3DCBA	3DACB
2BACD	2BACD	2CDAB	2CDAB	2DCAB	1ABCD	1ABCD
2DCBA	2CDAB	2CDBA	1ABCD	2DCBA	1ABDC	1ADCB
1ADBC	1BADC	2DBCA	1ACBD	1ACBD	1ADCB	1BACD
1BCDA	1CADB	1ABCD	1BACD	1ACDB	1BACD	1BDAC
1CADB	1CBDA	1BACD	1CDBA	1ADCB	1CDAB	1CADB
1CBAD	1CDBA	1CABD	1DBAC	1BADC	1CDBA	1CDBA
1DACB	1DABC	1CADB	1DBCA	1CADB	1DBAC	1DBAC
1DBCA	1DACB	1DABC		1CCAB		1DCBA
	1DBCA	1DACB		1CDBA		
				1CDBA		
				1DBAC		
				1CDAB		
				1DBCA		
13	14	14	12	18	12	13
Total Number of Different Orderings						

Table 8. Summary of the preference data from each evaluation method (Voice/Text Fanatics)

FI	PT	PQ	MQ	PSSUQ	AHP	REG
4DCAB	3CDBA	5DCBA	3ADBC	5CDBA	4CDBA	4CDBA
3CDAB	3DABC	4CDBA	3CDAB	4ADBC	3ADBC	3DCBA
2BCDA	3DCAB	3DACB	3CDBA	3CDAB	3DCAB	2ADCB
2CADB	2CADB	2CADB	3DCBA	2CADB	3DCBA	2BDAC
2DACB	2CDAB	2DABC	2ACDB	2DACB	2ACDB	2CADB
1ABCD	2DACB	2DCAB	2DABC	2DCAB	2DABC	2DACB
1ABDC	2DBCA	1ABDC	2DACB	1ABDC	2DACB	2DCAB
1ACDB	2DCBA	1ACDB	2DCAB	1ADCB	1ADCB	1BCAD
1ADBC	1ABDC	1ADBC	1BDCA	1BCDA	1BCDA	1BDCA
1ADCB	1ACDB	1ADCB	1CADB	1DABC	1BDCA	1CABD
1BDAC	1ADBC	1BDCA	1CBDA	1DBCA	1CADB	1CBDA
1BDCA	1BCDA	1CDAB	1ACDB	1BCDA	1CDAB	1CDAB
1CBAD	1BDCA					1DACB
1CBDA						1DBCA
1DABC						
1DBAC						
16	13	12	12	12	12	14
Total Number of Different Orderings						

From the summary of preference data, preference proportions between pairs were obtained, and accounted only for preferences between two phones. The preference proportion of AB is defined by the number of AB orderings (A before B in their preference ordering) divided by the total number of observations. If the proportion was greater than 0.5, phone A received the majority preference over phone B. According to the well-known social choice criterion by Condorcet (1785), a candidate should win if it is preferred to each rival candidate by a majority of voters. The preference proportion of each pair was shown for Minimalists (Table 9). From the table, phone D was favored over the other phones, while phone B was not favored over any other phone. This result was identical in both the Minimalists and Voice/Text Fanatics. In summary, preference proportions of AB, CA, CB, DA, DB, and DC are greater than 0.5 for both user groups.

Table 9. Preference proportion between pairs of phones by Minimalists

	AB	CA	CB	DA	DB	DC
FI	14/24	14/24	19/24	13/24	20/24	13/24
PT	13/24	14/24	19/24	15/24	20/24	14/24
PQ	16/24	15/24	19/24	15/24	21/24	13/24
MQ	15/24	17/24	20/24	16/24	21/24	13/24
PSSUQ	*11/24	14/24	17/24	16/24	20/24	13/24
AHP	15/24	16/24	17/24	17/24	21/24	16/24
REG	19/24	13/24	17/24	20/24	21/24	13/24
Mean	14.71/24	14.57/24	18.29/24	16.00/24	20.57/24	13.57/24

* Since 11 is less than 13, B is preferable over A (BA instead of AB) by PSSUQ.

Table 10. Preference proportion between pairs of phones by Voice/Text Fanatics

	AB	CA	CB	DA	DB	DC
FI	17/24	14/24	15/24	16/24	16/24	13/24
PT	15/24	16/24	15/24	19/24	21/24	15/24
PQ	14/24	15/24	19/24	18/24	22/24	16/24
MPUQ	16/24	14/24	18/24	17/24	22/24	13/24
PSSUQ	16/24	15/24	15/24	16/24	21/24	*12/24
AHP	15/24	14/24	17/24	17/24	22/24	15/24
REG	11/24	17/24	18/24	18/24	20/24	14/24
Mean	14.86/24	15.00/24	17.29/24	16.71/24	20.57/24	14.14/24

*12 indicates CD or DC is a tie by PSSUQ as underlined in **Error! Reference source not found.**

Based on the mean ranking, median, greatest number of 1st rank, least number of 4th rank, and Condorcet criteria, the winner was determined by each evaluation method for each user group (Table 11 and Table 12).

Table 11. Winner selection methods and results for Minimalists

Evaluation Methods	Methods to Select First Preference				Condorcet Winner
	Mean Rank	Median	Greatest # of 1 st Rank	Smallest # of 4 th Rank	
FI	D	C, D	D	C	D
PT	D	C, D	D	C, D	D
PQ	D	C, D	D	C	D
PSSUQ	D	C, D	D	C	D
MQ	D	D	D	C	D
AHP	D	D	D	C, D	D
REG	D	C, D	D	D	D

Table 12. Winner selection methods and results for Voice/Text Fanatics

Evaluation Methods	Methods to Select First Preference				Condorcet Winner
	Mean Rank	Median	Greatest # of 1 st Rank	Smallest # of 4 th Rank	
FI	D	C, D	D	D	D
PT	D	D	D	D	D
PQ	D	D	D	D	D
PSSUQ	D	C, D	C	D	D
MQ	D	C, D	D	D	D
AHP	D	C, D	D	D	D
REG	D	C, D	D	A, C, D	D

All the winner selections above were based on descriptive statistics rather than on a statistical test using significance level.

Friedman Test

To illustrate and interpret the ranked data effectively, a contingency table showing the frequency of ranks from each treatment in each cell was developed. For example, Table 13 shows the contingency table from the first set of ranked data (FI).

Table 13. Rankings of the four phones based on first impression

Phone	Rank				Total
	1	2	3	4	
A	7	4	6	7	24
B	3	2	6	13	24
C	5	11	7	1	24
D	9	7	5	3	24
Total	24	24	24	24	96

The important question is whether there is a significant difference between the phones in terms of ranking. Various test statistics are used to examine differences between treatments based on

ranked data. One of the popular tests is the Friedman test, which uses the sum of the overall responses of the ranks assigned to each treatment (phone). The null hypothesis is that there is no difference between the treatments. For the data set of FI, it was found that there were significant differences among the treatments (Friedman statistic $R = 11.35$, $p < 0.01$).

For further analysis of the significant difference in each pair, post hoc paired comparisons using unit normal distribution were performed. There were significant differences between phones B and C, and between phones B and D ($p < 0.05$), while all the other pairs showed no significant differences ($p > 0.05$). The summary of significant preference for all seven methods was provided for Minimalists (Table 14) and for Voice/Text Fanatics (Table 15).

Table 14. Summary of significant findings from Friedman test for Minimalist

Ranked Data	*Significant Preference ($p < 0.05$)
FI	DB, CB, AB
PT	DB, CB, DA
PQ	DB, CB, AB, DA
MQ	DB, CB, DA
PSSUQ	DB, DA, CB
AHP	DB, DA, CB
REG	DB, CB, AB

* In the order of larger rank sum difference by post hoc paired comparison using unit normal distribution

Table 15. Summary of significant findings from Friedman test for Voice/Text Fanatics

Ranked Data	Significant Preference ($p < 0.05$)
FI	None
PT	DB, DA, CB
PQ	DB, DA, CB
MQ	DB, CB, DA, AB
PSSUQ	DB, CB, DA
AHP	DB, CB, DA
REG	DB, DA, CB, CA

Important Usability Dimensions

According to the result of comparative evaluation, it is clear that phone D was the best phone in terms of usability and phone B was the worst. The mean scores from the MPUQ of each phone on each factor group were shown (Figure 5 and Figure 6). The EAMC score of phone B was significantly lower than that of the other phones ($p = 0.0006$). There was no significant difference in the scores of the other factors across phones.

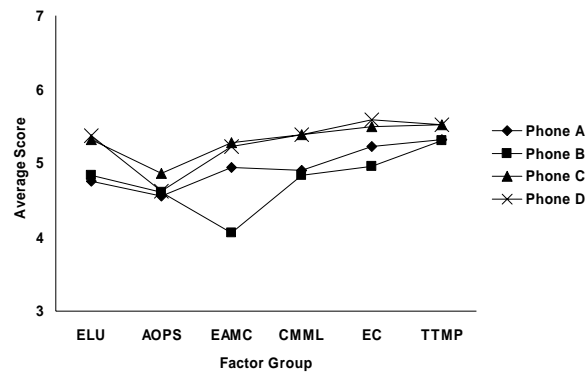


Figure 5. Mean scores on each factor of MPUQ for Minimalists

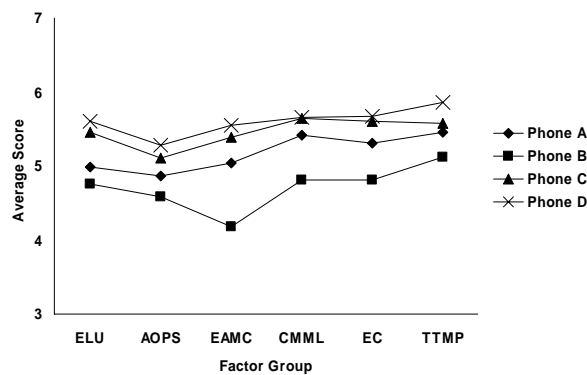


Figure 6. Mean scores on each factor of MPUQ for Voice/Text Fanatics

One interesting point relating to Voice/Text Fanatics was that the mean scores are in the order of phone D, C, A, and B for every factor. Also, there were significant differences in the scores of each factor across phones, except for TTMP: ELU ($p=0.0015$), AOPS ($p=0.0218$), EAMC ($p<0.0001$), CMML ($p=0.0013$), EC ($p=0.0003$), and TTMP ($p=0.0575$).

The trend showed that factor TTMP received a relatively higher score than others. Thus, it is plausible that most users simply do not find it challenging to perform typical tasks of using mobile phones, such as making and receiving phone calls, using the phonebook, checking call history and voice mail, and sending and receiving text messages.

From the pairwise comparisons by AHP, the important usability dimensions were identified for each user group (Ryu, Babski-Reeves, Smith-Jackson, & Nussbaum, 2007). Efficiency was most important to Minimalists and effectiveness was most important to the Voice/Text Fanatics. EAMC was the least important for both user groups. TTMP was the most important for the Voice/Text Fanatics, while EC was the most important for Minimalists.

This result is comparable to the result from REG, since EAMC is the greatest contributing factor for Minimalists and ELU, AOPS, and TTMP are the ones for Voice/Text Fanatics. Table 16 summarizes the comparable findings. As shown, there was no commonly significant usability dimension for Minimalists identified by both the AHP and REG. TTMP was the influential factor for Voice/Text Fanatics identified by the both methods.

Table 16. Decisive usability dimensions for each user group identified by the AHP and REG method

	AHP	REG
Minimalist	Efficiency and control	Emotional aspect and multimedia capabilities Ease of learning and use
Voice/Text Fanatics	Typical tasks for mobile phones	Assistance with operation and problem solving Typical tasks for mobile phones

Discussion

Implication of Each Evaluation Method

There were seven different methods for the comparative usability evaluation performed:

1. first-impression ranking (FI)
2. post-training ranking (PT)
3. post-questionnaire ranking (PQ)
4. ranking from the mean score of the MPUQ (MQ)
5. ranking from the mean score of the PSSUQ (PSSUQ)
6. ranking from the MPUQ model using AHP (AHP)
7. ranking from the regression model of MPUQ (REG)

FI allowed the participants to explore the phones briefly (less than 2 minutes), so that it might have been hard for them to grasp the context of usability of each phone. Thus, the decision of ordering each phone in terms of inclination to own one could be mostly based on the appearance of the phone design. In other words, the decision could rely heavily on affective and emotional aspects of the phones, which were the aspects the participants could most readily perceive in a brief time.

After the training session (PT; post-training), participants should have gained the context of usability by performing the predetermined tasks. Nineteen of each 24 Minimalists and 24 Voice/Text Fanatics changed their FI rank following the training. It could be inferred that the gaining the context of usability of each phone could affect decision-making. PT was used as the dependent variable to be predicted by REG, since the decision-making activity of PT would be the most analogous to the actual purchasing behavior, referred to as a descriptive model.

After answering the MPUQ and PSSUQ (PQ, post-questionnaire), six of 24 Minimalists changed their minds to re-order their rank orderings, and 11 of 24 Voice/Text Fanatics did. This means that the activity of answering usability questionnaires may have affected the decision-making process of the participants. The usability questionnaires played the role of enhancing the users' conceptualization of the context of usability and aiding users in making decisions. This finding is analogous to that of the developers of SUMI, who indicated that the activity of answering SUMI improves novice users' ability to specify design recommendations (Kirakowski, 1996). Thus, the activity of answering a usability questionnaire not only improves users' ability to provide specific design recommendations, but also affects users' decision-making process for comparative evaluation.

Usability and Actual Purchase

Participants were asked to determine phone rankings based on the preference by assuming all other factors, such as price and promotions, were identical. Since MQ, PSSUQ, AHP, and REG methods determined the ranks based on the scores from usability questionnaires, the decisions were not directly related to the intent of actual purchase. There has been little research on the relationship between usability and actual purchase of phones. According to the result of this study, performing the typical tasks of phones (PT) as well as answering the usability questionnaire (PQ) could influence the decision to select and purchase a phone.

Limitations

The superiority of phone D in terms of usability over the other phones may have interfered with the result of this study. Phone D was designed from extensive usability studies from a multi-year project performed by a Virginia Tech research team.

In this study, PT was set up as the dependent measure to develop regression models from the perspective that the decision by PT would be the closest decision of consumer's typical behavior. Thus, the correlation values of other methods with PT were investigated to determine which method is the best predictor of PT. However, it is difficult to argue that PT is the closest to the true value we want to predict.

Another limitation could be the population of users used in this research. Most of the participants were young college students. Because it was expected beforehand that the participant population would be limited to college students, the mobile user categorization (IDC, 2003) was applied to distinguish user profiles other than typical characteristics such as age, gender, and experience of usage. Thus, the results of this research would only be valid with the population of young college students representing each user group.

There were variations across the methods and models in terms of the number of orderings, preference proportions, and methods to select a first preference, while the mean ranking data was not much different across the methods and models. Thus, the study provides a useful insight into how users make different decisions through different evaluation methods.

Conclusion

In this paper, a case study of comparative usability evaluation was performed using various subjective evaluation methods, including MPUQ. The findings revealed that phone D, which was designed based on outcomes of several usability studies, was preferred across user groups. Although the mean rankings of the four phones were not significantly different across the evaluation methods, there were variations across the methods in terms of the number of rank orderings, preference proportions, and methods to select their initial preference.

Future Work

Since more than 70% of mobile users who participated in Ryu and Smith-Jackson (2006) were self-defined as either Minimalists or Voice/Text Fanatics, the development of the decision-making models and comparative evaluation in this paper were constrained to only these two user groups. Assuming that the other two users groups (i.e., Display Mavens and Mobile Elites) may have unique characteristics of usage and purchasing behavior, the studies with similarly large numbers of users from those two user groups would be beneficial to mobile device manufacturers.

One of the interesting findings of the current research was that the activity of answering usability questionnaires could be effective in changing the intentions to purchase. Although numerous usability studies of consumer phones have been conducted, few studies have investigated the direct relationship between usability and actual purchasing behavior. To establish the value of mobile phone design enhancement based on usability studies, extensive research to determine the relationship would be a promising direction for future research.

Practitioner's Take Away

The following important usability dimensions and items for mobile phone usability were based on MPUQ (Mobile Phone Usability Questionnaire) results using several usability evaluation methods (Ryu & Smith-Jackson, 2006):

- EC (Efficiency and Control) for Minimalists
- TTMP (Typical Task for Mobile Phones) for Voice/Text Fanatics

Thus, if usability practitioners want to employ a short list of questions to compare mobile phones for each user group, the questions from EC (9 questions) and TTMP (7 questions) of MPUQ could be selected as appropriate.

Also, the activity of answering a usability questionnaire has the following effects:

- improves users' ability to provide specific design recommendations (Kirakowski, 1996)
- affects users' decision-making process for comparative evaluation.

The results and outcomes of this paper were restricted to two major mobile user groups, Minimalists and Voice/Text Fanatics.

References

- Bell, D. E., Raiffa, H., & Tversky, A. (1988a). *Decision-making: Descriptive, Normative, and Prescriptive Interactions*. Cambridge: Cambridge University Press.
- Bell, D. E., Raiffa, H., & Tversky, A. (1988b). Descriptive, normative, and prescriptive interactions in decision-making. In D. E. Bell, H. Raiffa & A. Tversky (Eds.), *Decision-making: Descriptive, Normative, and Prescriptive Interactions*. Cambridge: Cambridge University Press.
- Condorcet, M. J. (1785). *Essai sur l'application de l'analyse a la probabilité des décisions rendues a la pluralité des voix*. Paris.
- Dillon, S. M. (1998). *Descriptive Decision-making: Comparing Theory with Practice*. In Proceedings of 33rd ORSNZ Conference, University of Auckland, New Zealand.
- Harper, P. D., & Norman, K. L. (1993). *Improving user satisfaction: The questionnaire for user interaction satisfaction version 5.5*. In Proceedings of The 1st Annual Mid-Atlantic Human Factors Conference, Virginia Beach, VA.
- IDC. (2003). *Exploring Usage Models in Mobility: A Cluster Analysis of Mobile Users*. (No. IDC #30358): International Data Corporation.
- Kirakowski, J. (1996). The software usability measurement inventory: Background and usage. In P. W. Jordan, B. Thomas, B. A. Weerdmeester & I. L. McClelland (Eds.), *Usability Evaluation In Industry* (pp. 169-178). London: Taylor & Francis.
- Kirakowski, J., & Corbett, M. (1993). SUMI: The software usability measurement inventory. *British Journal of Educational Technology*, 24(3), 210-212.
- Lewis, J. R. (1995). IBM computer usability satisfaction questionnaire: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1), 57-78.
- Lin, H. X., Choong, Y.-Y., & Salvendy, G. (1997). A Proposed Index of Usability: A Method for Comparing the Relative Usability of Different Software Systems. *Behaviour & Information Technology*, 16(4/5), 267-278.
- Nunnally, J. C. (1978). *Psychometric Theory*. New York: McGraw-Hill.
- Payne, J. W. (1976). Task complexity and contingent processing in decision-making: An information search and protocol analysis. *Organizational Behavior and Human Performance*, 16, 366-387.
- Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The Adaptive Decision Maker*. Cambridge: Cambridge University Press.

Ryu, Y. S., Babski-Reeves, K., Smith-Jackson, T. L., & Nussbaum, M. A. (2007). *Enhancing the Mobile Phone Usability Questionnaire (MPUQ) with a Multi-criteria Decision-making Method*. In Proceedings of Human Computer Interaction International 2007, Beijing, China.

Ryu, Y. S., & Smith-Jackson, T. L. (2006). Reliability and Validity of the Mobile Phone Usability Questionnaire (MPUQ). *Journal of Usability Studies*, 2(1), 39-53.

Schoemaker, P. J. H. (1980). *Experiments On Decisions Under Risk: The Expected Utility Theorem*. Boston, MA: Martinus Nijhoff Publishing.

Taplin, R. H. (1997). The statistical analysis of preference data. *Applied Statistics*, 46(4), 493-512.

About the Authors



Young Sam Ryu

is an Assistant Professor of Ingram School of Engineering at Texas State University-San Marcos. His research area includes human computer interaction, usability engineering, and safety. He received his Ph.D. (2005) from the Grado Department of Industrial and Systems Engineering at Virginia Tech.



Kari Babski-Reeves

is an Assistant Professor at Mississippi State University. Her background in Industrial and Systems Engineering is supplemented with training in kinesiology and management. Her research interests are applications of thermal imaging to ergonomics research, operator fatigue, training systems, and work related musculoskeletal disorders.



Tonya L. Smith-Jackson

is an Associate Professor in the Grado Department of Industrial and Systems Engineering (ISE), where she teaches courses in human factors engineering, cognition, usability, and safety systems. She is director and co-director of 3 laboratories, and is also the director of the Human Factors Engineering and Ergonomics Center. Her usability work focuses on software and hardware interfaces for mobile phones, requirements elicitation and application to design, and in finding innovative and practical ways to elicit and assess latent factors in usability.



Maury A. Nussbaum

received an M.S. (1989) in Bioengineering and a Ph.D. (1994) in Industrial and Operations Engineering from The University of Michigan. He is currently a Professor of Grado Department of Industrial and Systems Engineering at Virginia Tech. His research interests are occupational biomechanics and ergonomics, aging, postural control, and consumer product design and evaluation.