December 2007

Measuring Mobile Device Usability as a Second Order Construct in Mobile Information Systems

Andrew Urbaczewski

Matti Koivisto
University of Michigan - Dearborn

MEASURING MOBILE DEVICE USABILITY AS A SECOND ORDER CONSTRUCT IN MOBILE INFORMATION SYSTEMS

Andrew Urbaczewski
University of Michigan – Dearborn
aurbacze@umd.umich.edu

Matti Koivisto
Mikkeli University of Applied Sciences
Matti.Koivisto@mikkeliamk.fi

ABSTRACT:

There are many existing models of usability and instruments to measure it, such as the Administrative Scenario Questionnaire (ASQ) and the Satisfaction Usability Scale (SUS), and the Network Satisfaction Scale (NET). Performance metrics also exist to measure efficiency and effectiveness. However, none of these instruments have been used, separately or together, to measure the overall usability of mobile devices used in m-commerce. A study was conducted to test the usability of mobile devices, and multiple existing metrics were used in order to get an overall view of usability as a second order construct. Confirmatory Factor Analysis found that effectiveness and NET dropped out of the model's analysis, but a respecified model confirmed that usability was a second order construct, predicted by efficiency, SUS, and ASQ.

Keywords: Mobile Information Systems, Construct Development, Usability

Introduction

Mobile telephony has undoubtedly changed the way that individuals function during the 21st century. First developed as a voice technology for the rich, mobile telephony has become almost ubiquitous in most of the developed world today. Information providers and organizations have sought to identify ways to deliver information to these ubiquitous terminals.

While the mobile phone has evolved in its size and network used, it's basic shape and input mechanisms have remained the same. These are modeled after the handset for the normal wired telephone since the speaker and the microphone have been placed together in the mid 20th century. The input methodology has also remained the same since the introduction of touch-tone dialing in the 1960s – 12 buttons arranged in a 3*4 grid, with the 10 Arabic numerals and * and # buttons.

Modern phones are fine for voice conversations, but the device that works ideally for one purpose may be ill-suited for others. In the approximately 20 years of the mobile phone industry, personal computers have become ubiquitous in society. While the form factor of computers has also gotten smaller, they are not anywhere near as small as mobile phones. In many ways, the "mobile personal computer", i.e., the laptop, has gotten bigger in the last couple of years, due to falling prices of high-quality TFT glass for the LCD panels used for the display.

Though today's mobile devices have advances in display size, and colors, there is still the physical limitation of making a device that will fit the user's hand AND provide a large display and uncomplicated user interface. Many new "smartphones" provide larger displays and full QWERTY keyboards, but yet have little acceptance outside of the executive business community. Even so, a focus on usability has been necessary to ensure the maximum efficiency of investments in hardware and software in the delivery of new services designed for the smaller devices, such as sales force automation (SFA) applications, limited Web browsing, and full e-mail services.

There are many definitions of usability, and many tools to measure that usability. The aim of this paper is to examine different tools used to measure usability, administer them in a mobile device context, and then compare them to gain a greater understanding of the usability phenomenon. Our experiment measured mobile usability in three different contexts, and examine if these devices are truly measuring the same phenomenon, or different phenomena that make up usability.

Usability

Usability is critical to the success of mobile devices and acceptance of mobile technology in general. But what is usability and how should it be measured? Researchers agree that usability involves many mutually dependent dimensions (Holcomb and Tharp 1991, Nielsen 1993), but many different classifications of these dimensions exist. Perhaps the most widely accepted definition of usability comes from ISO 9241, which defines usability as the effectiveness, efficiency, and satisfaction with which specified users can achieve specified goals in particular environment (ISO 9241-11, 1998).

Shackel (1990) has stated that for a system to be usable it has to achieve defined levels on the following attributes: effectiveness, learnability, flexibility and attitude. Nielsen (1993) instead associates usability with five attributes, which are learnability, efficiency, memorability, lack of errors (or accuracy) and satisfaction. Nielsen's usability attributes match Shneidermann's (1997) five measurable human factors goals of user interface design which are: time to learn, speed of performance, rate of errors by user, retention over time and subjective satisfaction. Differences in these perspectices were recognized by Keinonen (1998) and are listed in Figure 1.

Usability measures	**Usability point of view**	**Usability objectives**
- Errors - Time - Ratings	- Utility - Efficiency - Satisfaction	- Experienced user performance - Learnability - Relearnability

Figure 1. Measurements, objectives and views on usability (Keinonen 1998)

The model recognizes three measures of usability: the number of errors, performance time, and answers on a rating instrument; three design objectives: the experienced user´s performance, learnability by novice users, and relearnability or retention over time by casual users; and three views – the process output view i.e. utility, the resource usage view i.e. efficiency, and the user´s subjective view.

Measuring Usability

As difficult as it is to define usability and its dimensions, it is even more challenging to measure it. Many scholars (Bevan 1995, Larson 2002) agree that there are two types of usability measurements: preference and performance measurements. In performance measurements we try to collect objective metrics of the system performance. In preference measurements we are interested in user subjective preferences and opinion data. Both measurement types are discussed below.

Performance Measurements

System performance can be measured in many ways, but ISO's usability definition shows that usability related performance can be divided into two concepts: efficiency and effectiveness. Measures of efficiency relate the effectiveness achieved to the expenditure of resources. From a user's point of view the time and effort used for the task are resources he or she consumes. Measures of effectiveness relate instead the goals or sub-goals of using the system to the accuracy and completeness with which these goals can be achieved (Bevan and Macleod 1994).

In text entry evaluations efficiency is usually measured as input speed or throughput. Speed is usually calculated in characters per second or even more often as words per minute (wpm). These metrics are actually identical because the definition of a word for this purpose is five characters, including spaces or any other characters in the text.

The effectiveness of an input method is normally analyzed as accuracy. If calculations of entry speed are straightforward, accuracy is another matter. Even the intuitively simple measure "percent errors" is problematic, and differing methods like Levenshtein Minimum String Distance (Soukoreff and MacKenzie 2001), Character Level Error Rate (MacKenzie and Soukoreff 2002), and Word Error Rate (Wang et al. 2003) are used.

Efforts are underway to streamline and standardize text entry experiments. In particular, Soukoreff and MacKenzie have made an important contribution in this field. Contemporary Soukoreff-MacKenzie (2004) accuracy metrics are based on delineating participants´ keystrokes into four classes:

- Correct (C) keystrokes – alphanumeric keystrokes that are not errors,
- Incorrect and Not Fixed (INF) keystrokes – errors that go unnoticed and appear in the transcribed text
- Incorrect but Fixed (IF) keystrokes – erroneous keystrokes in the input stream that are later corrected, and,
- Fixes (F) – the keystrokes that perform the corrections (i.e. delete, backspace, cursor movement)

Based on this classification several statistics can be easily calculated, for example:

Total Error Rate = (INF + IF) / (C+INF+IF) * 100%
Not Corrected Error Rate = INF / (C+INF+IF) * 100%
Corrected Error Rate = IF / (C+INF+IF) * 100%

Preference measurements

If performance of the system is measured with objective data, preferences are subjective by nature. In ISO's usability definition, preferences measures are related to the basic usability concept of satisfaction. Measures of satisfaction describe the perceived usability of the overall system or some specific aspects of the system (Bevan and Macleod 1994). There are two major approaches for the measurement of user preferences: opinion polls and customer satisfaction surveys (Noll 1999).

The standard tool for analyzing users´ perception of the usability of a product is a usability questionnaire. A number of usability questionnaires have been developed during the last decades. Each of these instruments seeks to measure the usability construct in and of itself. Some questionnaires like SUS (System Usability Scale) (Brooke 1986) gives only one single value about the usability as a whole and others like SUMI (Software Usability Measurement Inventory) (Kirakowsky

1996) provide multiple scores. Also the number of questions in questionnaires varies. For example in After-Scenario Questionnaire (ASQ) (Lewis 1995), which is concentrating on user satisfaction, uses only three questions and questionnaires with broader scope like SUMI or SUS have more questions. In addition to these standard questionnaires some researchers have used their own set of questions in order to analyze some aspect of usability in more details. For example Koivisto and Urbaczewski (2004) used four questions in their network performance questionnaire (NET) to analyze the effect of network speed to the user-perceived quality of service of mobile Internet. Table 1 shows the questions from three different usability questionnaires.

Table 1. Questions in SUS, ASQ and NET Questionnaires

Questions in SUS questionnaire (Likert scale 1-5)	
1	I think I would like to use this system frequently.
2	I found the system unnecessarily complex.
3	I thought the system was easy to use.
4	I think that I would need the support of a technical person to be able to use this system.
5	I found the various functions in this system were well integrated.
6	I thought there was too much inconsistency in this system.
7	I would imagine that most people would learn to use this system very quickly.
8	I found the system very cumbersome to use.
9	I felt very confident using the system.
10	I need to learn a lot of things before I could get going with this system

Questions in ASQ questionnaire (Likert scale 1-7)	
1	Overall, I am satisfied with the ease of completing the tasks in this scenario
2	Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario
3	Overall, I am satisfied with the support information (online-line help, messages, documentation) when completing the tasks

Questions in NET questionnaire (Likert scale 1-5)	
1	Overall, I am satisfied with Quality of Service in the connection establishment
2	Overall, I am satisfied with the Quality of Service in data transfer
3	Overall, I am satisfied with the Quality of Service in the connection release
4	Overall, I am satisfied with the general Quality of Service of the network connection

If there was a clear understanding of usability, it stands to reason that device makers would not be having nearly the difficulty today in making mobile devices more usable. Therefore, we pose that usability is really a second-order construct, made up of bits and pieces of previously developed instruments but not wholly captured by any of them. We further posit that these measures are reflective of usability but that each of them add a certain dimension to the usability scenario:

H1: Increasing levels of usability in mobile devices will be reflected by increasing levels of efficiency.

H2: Increasing levels of usability in mobile devices will be reflected by increasing levels of effectiveness.

H3: Increasing levels of usability in mobile devices will be reflected by increasing SUS scores.

H4: Increasing levels of usability in mobile devices will be reflected by increasing ASQ scores.

H5: Increasing levels of usability in mobile devices will be reflected by increasing NET scores.

Methodology

The aim of our study is to identify an overall measurement of mobile device usability. To do this, we gathered as much data as possible using as many instruments as possible. We gathered information on the efficiency, effectiveness, SUS, ASQ, and NET dimensions when a group of test users wrote email messages with three different input methods. We used confirmatory factor analysis to measure the paths such that significant links could be tested.

In our experiment the input methods were stylus pen, multi tap, and reduced QWERTY keyboard and the device used was a PDA (a Compaq iPaq 3870 PDA with IEEE 802.11(b) WLAN connections). We also used three different message lengths (21, 63 and 197 characters) to study possible effects related to the number of characters, and the three different input methods were to measure the phenomena across multiple methods. The purpose was *not* to compare one input device against another to see which one was "best."

To be able to collect the required information for performance metrics calculations (like presses of backspace etc.) we bypassed the operating system's standard input methods and wrote the user interface totally with Macromedia Flash. For example the pressing a letter ´a´ in a keyboard did not directly enter a letter to the text field in the user interface. Instead an Action script connected to the "On Release" event of invisible button was called and the code added a letter to the display. For the same reason we were not using the operating system's soft keyboard but created our own soft keyboard (see Figure 2).

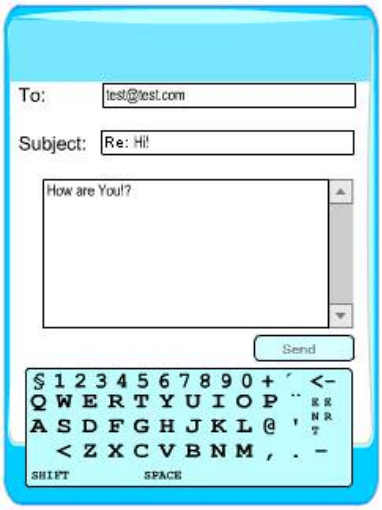

Figure 2. Our soft keyboard layout.

Even though multi-tap is a widely used input method in mobile phones it is not a standard feature in PDA devices. We implemented the multi-tap input method for a PDA with reduced QWERTY keyboard by re-labeling the used keys and covering the unused ones. Figure 3 shows devices used with the different input methods.

Figure 3: Devices used in the experiment.

The messages we used are shown in Table 2. It should be noted that in two messages (called a standard and a start of dialogue message) users filled in three fields (receiver's address, topic and message), and in the reply message only one field (message).

Table 2. Messages used in the experiment.

Type	Field	Content
Reply message	Address Topic Message	- - tuesday is ok see you
Standard message	Address Topic Message	sara@rock.net message the quick brown fox jumps over the lazy dog
Start of dialogue message	Address Topic Message	joe@mail.com hi joe how are you want to meet tonight want to go to the movie with sue and me what show do you want to see we are meeting in front of the theatre at eight let me know if we should wait

The data collection took place in a laboratory study in which undergraduate students of a large polytechnic school in Finland wrote email messages with three different input methods. Eighty-seven subjects (64 male, 23 female) participated in the study. Their average age was 24.6 years, with 9.09 years of computer usage, 5.77 years of mobile device usage and text message (SMS) experience. Three different input methods were used to reduce any claim that the findings were limited to one type of input methodology. Because each participant used the three different input methods, there were 261 total cases. Four cases were removed from the analysis because the test failed for some reason (e.g. the mobile phone of the test user rang during the test). We employed a Latin square technique to avoid a learning effect tainting the subjects and the results.

Because the instruments use different scales, transformations were done on the data such that they would all be representative of a 1-5 scale. The performance metrics were also recoded onto a 1-5 scale, using equal proportions to represent different

transformed scores. For example, the highest numbers of words per minute were scored as 5 and the lowest numbers of words per minute were recoded as a 1.

Results

Test of the Measurement Model

In keeping with extant research, we adopted a two-step approach in which we first established a valid and reliable measurement model prior to testing the hypothesized second-order factor model (Anderson and Gerbing 1988). Specifically, using EQS 6.1, the 18 – item, 5 –construct measurement model, containing the constructs effectiveness, SUS, and ASQ, NET, and efficiency, was subjected to confirmatory factor analysis (CFA) to test for convergent and discriminant validity, and reliabilities (internal consistency) were tested using Cronbach's alpha.

The resulting measures of model fit for the CFA were: CFI = .957, Robust CFI = .960, NFI = .922, NNFI = .947, a chi-square of 209.089, 124 df, and no standardized residuals greater than the absolute value of 2. Thus, the overall fit of the model was deemed acceptable (Anderson and Gerbing 1982, 1988; Bozdogan 1987; Chin 1998; Hu and Bentler 1999). However, when the evidence for convergent validity was examined, it was noted that 1 of the 2 items (NCER) prespecified to load on the effectiveness construct did not obtain a standardized factor loading exceeding 0.5 and the test of its unstandardized coefficient was not significant (t-value was -.204).

In our next step in purifying the measurement model, we incorporated these results and respecified the measurement model by eliminating the effectiveness construct. We then performed a CFA on the respecified 16-item, 4-construct model.

As presented in Table 3, the resulting measures of fit for the respecified model were: CFI = .970, Robust CFI = .973, NFI = .935, NNFI = .962, a chi-square of 147.946, 93 df, and no standardized residuals greater than the absolute value of 2. Thus, the overall fit of the proposed measurement model with the data was deemed acceptable (Anderson and Gerbing 1982, 1988; Bozdogan 1987; Chin 1998; Hu and Bentler 1999). As evidence of convergent validity, each of the 16 items loaded on their prespecified constructs, all standardized factor loadings exceeded 0.5, and all tests of unstandardized coefficients were significant (t-values were between 6.888 and 12.809; see Table 2 for standardized estimates). Also, as shown in Table 3, all scales achieved Cronbach's alpha greater than .70 and therefore were deemed acceptably convergent (Nunnally 1978). To establish discriminant validity, a multivariate LaGrange multiplier (LM) test indicated no significant cross-loadings for measurement items with non-hypothesized constructs. Thus the measurement model was considered sufficiently reliable and valid (Anderson and Gerbing 1982, 1988; Chin 1998; Fornell and Larcker, 1981).

Table 3 - Model Construct Measures and Reliabilities Based on CFA

Construct	Measures	Standardized Parameter *	α
Efficiency			.753
WPM	Words per minute	.956	
KSPC	Keystrokes per Character	-.516	
SUS			
1	I think I would like to use this system frequently.	.764	.820
2	I found the system unnecessarily complex.	.610	
3	I thought the system was easy to use.	.829	
6	I thought there was too much inconsistency in this system.	.546	
7	I would imagine that most people would learn to use this system very quickly.	.546	
8	I found the system very cumbersome to use.	.511	
9	I felt very confident using the system.	.713	
ASQ			
1	Overall, I am satisfied with the ease of completing the tasks in this scenario	.828	.770
2	Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario	.785	
4	Overall, I am satisfied with the support information (online-line help, messages, documentation) when completing the tasks	.625	
NET			
1	Overall, I am satisfied with the Quality of Service in the connection establishment	.878	.909
2	Overall, I am satsifed with the Quality of Service in data transfer	.884	
3	Overall, I am satisfied with the Quality of Service in the connection release	.863	
4	Overall I am satisfied with the general Quality of Service of the network connection	.777	

Goodness of Fit Results: Chi-square of 147.946, df 93; CFI=.970
*** Significant at p<.01**

Test of the Second-Order Factor Model

After determining that the respecified measurement model was sufficiently valid and reliable, the second-order factor model presented in Figure 4 was tested. The purpose was to determine whether the 4 primary dimensions (efficiency, SUS, ASQ, and NET) can be viewed as appropriate indicators of IT Usability. As shown in Table 4, the results indicate that the model fits the data well (CFI = .971, Robust CFI = .973, NFI = .928, NNFI = .963, chi-square of 145.051, 92 df, and no standardized residuals greater than the absolute value of 2).

Table 4 presents the standardized parameters and t-values resulting from the testing and indicates that three of the 4 constructs – efficiency, SUS, and ASQ – are dimensions of the second-order factor IT usability. However, the results do not support the inclusion of NET as a dimension.

Table 4 - Tests of Hypothesized Relationships for Second-Order Factor Model

Hypothesis	Standardized Parameter Estimate*	t-values	Conclusion
H1: EFY directly influences IT usability.	.482	7.099	Supported
H2: SUS directly influences IT usability.	.928	9.345	Supported
H3: ASQ directly influences IT usability.	.958	13.352	Supported
H4: NET directly influences IT usability.	.078	1.080	Not Supported

Goodness of Fit Results: Chi-square of 145.051, df 92; CFI=.971.
***significant at p < .01**

Incorporating this finding, we respecified the second-order factor model in an attempt to increase the degree to which the second order factor model fits the data and, therefore, improve the validity of the conceptualization (Bentler and Chou 1987). Specifically, based on the empirical results, we elected to eliminate the path between IT usability and NET. The resulting measures of model fit for the respecified model were: CFI = .977, Robust CFI = .981, NFI = .942, NNFI = .966, chi-square of 72.006, 73 df, and no standardized residuals greater than the absolute value of 2. Thus, all measures of model fit improved with model respecification. As a next step, we evaluated the change in overall goodness of fit between the two models using the chi-square difference test (Bollen, 1989). The chi-square difference of the two models was 73.045, 47 df, $p < .05$, indicating there is significant difference in overall fit between the two models. Additionally, further scrutiny of the remaining hypothesized paths was keeping with previous model results (see Table 5 that presents the standardized parameters and t-values resulting from the testing). Therefore, we concluded that respecifying to a three construct structure did improve upon the initially proposed four-construct second-order factor conceptualization of IT usability.

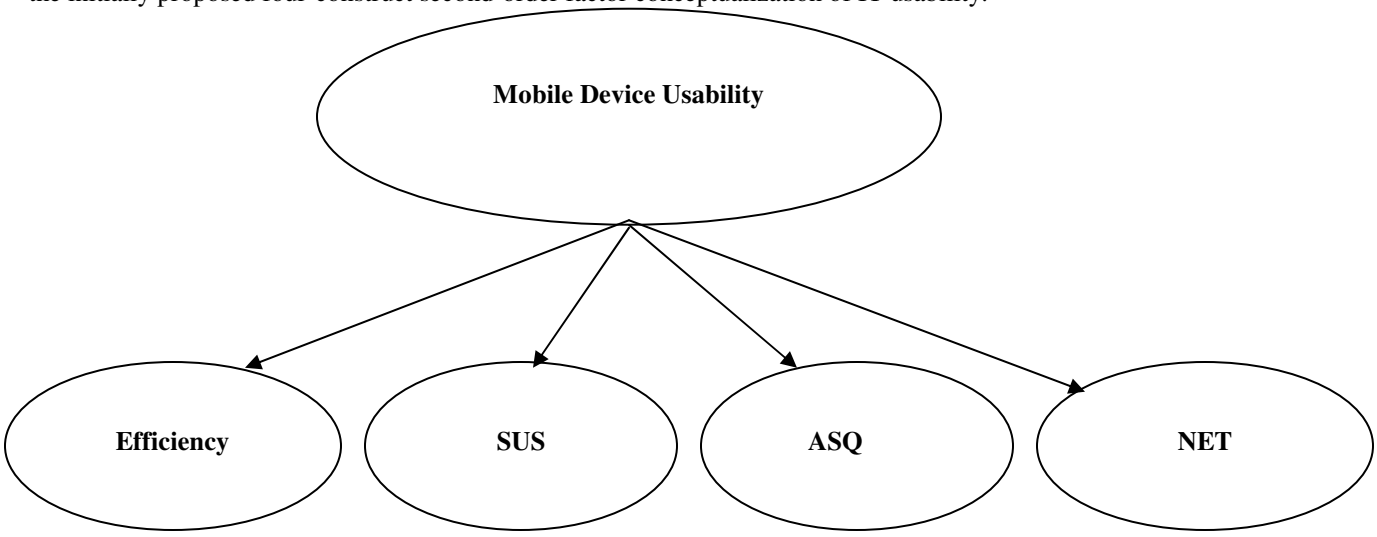

Figure 4 - Hypothesized Second-Order Factor Model

Table 5 - Tests of Hypothesized Relationships for
Respecified Second-Order Factor Model

Hypothesis	Standardized Parameter Estimate*	t-values	Conclusion
H1: EFY directly influences IT usability.	.476	7.475	Supported
H2: SUS directly influences IT usability.	.931	9.424	Supported
H3: ASQ directly influences IT usability.	.960	13.404	Supported

Goodness of Fit Results: Chi-square of 72.006, df 45; CFI=.977.
***significant at p < .01**

Discussion

This experiment to measure usability and compare metrics has resulted in many important findings. First of all, it confirms that usability is indeed a second order construct, made up of many smaller constructs. While the work of others has hinted this would be the case (e.g., Schackel 1990, Nielsen 1993, ISO 9241-11 1998), this is the first work to empirically measure the components of usability. We searched for components of effectiveness, efficiency, and satisfaction, as predicted in the ISO 9241 definition, and confirmed some findings while disconfirming others.

Effectiveness, as measured by CER and NCER, as proposed by Soukoreff and MacKenzie (2004), proved to not be a construct. NCER was the problem variable, not loading on the construct and having an insignificant t-score. The authors are unaware of any work that tests the validity of CER and NCER as accurate measures of usability, and this study brings Soukoreff and MacKenzie's propositions into question.

Furthermore, it appears that KSPC, long theorized as a measure of effectiveness (Soukoreff and MacKenzie 2004), is really a measure of efficiency. KSPC loaded with WPM on a single construct and was a significant contributor to the construct. It is possible this is a unique factor to the mobile world, which can rely on multi-tap devices to generate a character (e.g, pressing the "4" button three times to generate the "I" character.

Another finding of this study, somewhat surprisingly, was that SUS and ASQ measure different constructs. It was presumed that both instruments were proxies for the same usability construct. The confirmatory factor analysis showed that these are

indeed different constructs. A review of the purified measures seem to indicate that the SUS may be more related to the device characteristics and the ASQ may be more related to the actual task being completed.

Conclusions

This project did not seek to identify an "optimal" model of input or output. This was a project designed rather to study the phenomenon of usability, defining its subconstructs, and assisting others in identifying proper measures for usability. While ISO 9241 defines usability in general, this paper confirmed that usability for mobile devices is not radically different from other types of system usability. The major divide in this area was that KSPC loaded with efficiency. Further research should identify if this construct continues to hold with other devices, such as mobile phone inputs that use T9 predictive text input.

This paper also showed the user satisfaction areas are indeed quite important and should not be ignored. We identified that while there are many metrics that claim to measure "satisfaction," they may indeed be measuring slightly different constructs. If our suspicions about the nature of SUS and ASQ are true, that is, that SUS measures system elements and ASQ measures task elements, this would lead to additional research that relates mobile usability to the task-technology fit (Goodhue and Thompson 1995) stream of research. More studies should be conducted in this area.

There are also limitations to this study. It cannot be ignored that this study was conducted in one country using a sample of participants with relatively similar cultural and demographic dimensions. However, this sample is representative of the young people that use mobile phones, and that all of them were mobile phone users and 96% sent at least one SMS per day, we can say that these individuals were not novices in the mobile device world.

It is also possible that the email task that students were asked to complete created an artifact. Perhaps other tasks, such as mobile banking or news searches might have created different results. Furthermore, though we purposely created different tasks (though all involved email) and different input methods in order to reduce or eliminate the possibility that a phenomenon was device or task dependent, it is possible that the multi-tap was not realistic enough. We took great care in designing the multi-tap interface, but that it was not the actual mobile phone might have created some external validity concerns.

Mobile usability is a topic of growing importance. It does no good to create complicated and sophisticated mobile information systems that are unusable due to input or display constraints. Continued study of mobile usability will help systems analysts, hardware developers, and software developers to create real, usable, systems.

References

Anderson, J., and D. Gerbing. (1982). "Some Methods for Respecifying Measurement Models to Obtain Unidimensional Construct Measures." *Journal of Marketing Research 19*: 453-60.

Anderson, J., and D. Gerbing. (1988). "Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach." *Psychological Bulletin*, *103(May)*: 411-423.

Bentler, P. and C. Chou (1987). "Practical Issues In Structural Equation Modeling." *Sociological Methods and Research, 16(1):* 78-117.

Bevan , N. (1995). "Measuring Usability as Quality of Use." *Software Quality Journal 4:* 115-130.

Bevan, N. and N. Macleod. (1994). "Usability Measurement in Context." *Behaviour and Information Technology, 13:* 132-145.

Bollen, K. (1989). *Structural Equations with Latent Variables*, John Wiley & Sons, New York.

Bozdogan, H. 1987. "Model Slection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions." *Psychometrika 52(3)*: 345-370

Brooke, J. (1986) *System Usability Scale*, Digital Equipment Corporation

Chin, W. 1998. "Issues and Opinion on Structural Equation Modeling." *MIS Quarterly*, *22(1)* : 7-16.

Fornell, C., and B. Larcker (1981). "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error." *Journal of Marketing Research 18(1)*: 39-50.

Goodhue D. and R.L. Thompson. (1995). "Task-Technology Fit and Individual Performance." *MIS Quarterly 19(2):* 213-236.

Holcomb, R. and A. Tharp. (1991). "What Users Say about Software Usability." *International Journal of Human-Computer Interaction 3(1):* 49-78.

Hu, L., and P. Bentler (1999). "Cutoff Criteria for Fit Indexes In Covariance Structure Analysis: Conventional Criteria versus Alternatives." *Structural Equation Modeling 6(1)*: 1-55.

ISO 9241-11 (1998). *Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) – Part 11. Guidance on Usability*

Keinonen, T. (1998). "One-dimensional Usability - Influence of Usability on Consumers' Product Preference." UIAH publication A21. Helsinki 1998.

Kirakowsky, J. (1996). "The software usability measurement inventory: Background and usage." In Jordan, P. W.; Thomas, B; Weerdmeester, B. A. and McClelland, I. L. (eds.). *Usability evaluation in industry*. Taylor & Francis, London, 169-178

Koivisto, M. and A. Urbaczewski. (2004). "The Relationship Between Quality of Service Perceived and Delivered in Mobile Internet Communications." *Information Systems and e-Business Management 2(4)*: 309-323.

Larson, J. 2002. "The What, Why, and How of Usability Testing." *Speech Technology Magazine 7(5).*

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

MacKenzie, I. S., and R.W. Soukoreff. (2002). "A character-level error analysis technique for evaluating text entry methods." *Proceedings of the Second Nordic Conference on Human-Computer Interaction – NordiCHI 2002,* 241-244. New York: ACM

Nielsen, J. (1993). *Usability Engineering*, Boston MA, Academic Press.

Noll, J. (ed) (1999) *Quality of Service (QoS) Measures for Applications*, EURESCOM Project P921

Nunnally, J. C. (1978). *Psychometric Theory*, Second Edition, McGraw-Hill, New York.

Schakel, B. (1990). "Usability-context, framework, definition, design and evaluation." In B. Schakel & S. Richardson (Eds.), *Human factors for information usability*. Cambridge: Cambridge University Press. 21-37.

Shneiderman, B. (1997) *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 3rd Edition, Addison-Wesley.

Soukoreff W. and S. MacKenzie. (2001). "Measuring Errors in Text Entry Tasks: An Application of the Levenshtein String Distance Statistic." *ACM Conference on Human Factors in Computing Systems - CHI 2001*

Soukoreff, R. W., and I.S. MacKenzie. (2004). "Recent developments in text entry error rate measurements." *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems*, 1425-1428. New York: ACM

Wang Y., Acero, A. and C. Chelba. (2003). "Is Word Error Rate a Good Indicator for Spoken Language Understanding Accuracy?" *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding*. Virgin Islands, Dec, 2003.