AIS Transactions on Human-Computer Interaction

Volume 6 | Issue 4 Article 1

2014

Assessing Smartphone Ease of Use and Learning from the Perspective of Novice and Expert Users: Development and Illustration of Mobile Benchmark Tasks

Dennis F. Galletta
University of Pittsburgh - Main Campus, galletta@katz.pitt.edu

Brian K. Dunn
University of Oklahoma, bkdunn@ou.edu

Transactions on
Human-Computer Interaction

THCI

Original Research

Assessing Smartphone Ease of Use and Learning from the Perspective of Novice and Expert Users: Development and Illustration of Mobile Benchmark Tasks

Dennis F. Galletta
Katz Graduate School of Business
University of Pittsburgh
galletta@katz.pitt.edu

Brian Kimball Dunn
Price College of Business
University of Oklahoma
bkdunn@ou.edu

Abstract

Assessing usability of device types with novel function sets that are adopted by diverse user groups requires one to explore a variety of approaches. In this paper, we develop such an approach to assess usability of smartphone devices. Using a three-stage Delphi-method study, we identify sets of benchmark tasks that can be used to assess usability for various user types. These task sets enable one to evaluate smartphone platforms from two perspectives: ease of learning (for those unfamiliar with smartphone use) and ease of use (for experienced users). We then demonstrate an approach for using this task set by performing an exploratory study of both inexperienced smartphone users (using a convenience sample) and experienced users (using the keystroke model). Our exploration illustrates the methodology for using such a task set and, in so doing, reveals significant differences among the leading smartphone platforms between novice and expert users. As such, we provide some preliminary evidence that ease of use is indeed significantly different from ease of learning.

Keywords: Ease of Use, Usability, Learnability, Keystroke Model, Mobile, Mobile Operating Systems, Windows, Blackberry, Android, iPhone

Joe Valacich was the accepting Senior Editor.

Galletta, D. F., & Dunn, B. K. (2014). Assessing smartphone ease of use and learning from the perspective of novice and expert users: Development and illustration of mobile benchmark tasks, *THCI*, *6*(4), 74-91.

1. Introduction

New types of multi-function computing devices with novel interface types and standards have risen in prominence over the last several years. In particular, smartphones have become a ubiquitous phenomenon. As of 2014, 71 percent of all Americans own smartphones (Nielsen, 2014a), and 85 percent of all mobile phone owners opt to upgrade to smartphones (Nielsen, 2014b). Further, the IDC has predicted that, by 2017, 70.5 percent of all connected devices in the world will be smartphones (Columbus, 2013).

With such devices becoming the de facto standard for both interpersonal communication and online connectivity, it has become increasingly important for researchers and practitioners alike to understand the dynamics behind their adoption and use. Most smartphone interfaces differ significantly from those of personal computers (PCs), largely due to their small size and the rarity of PC touchscreens. Further, Norman and Nielsen (2010) report that designers of smartphone and tablet interfaces have ignored years of accepted interface design standards by introducing designs that are confusing and foreign to users. Interestingly, as touchscreen PCs proliferate, the trend toward making computer interfaces resemble tablet interfaces could make operating a PC confusing, too, which would require many people to re-learn how to use them.

Instead of requiring mouse movements and keystrokes, smartphones rely on a new set of inputs such as swipes, pinches, voice-to-text transcriptions, and double-taps. Further, smartphones proffer a set of functionalities distinct from those available through traditional PCs and traditional cell phones; PCs are not often used for on-the-go photography, for instance, and traditional cell phone functionality is not extensible through downloading and installing applications.

Given smartphones' important differentiators from past devices and the diffusion of manufacturers and operating systems available for them, understanding user perceptions of smartphone usability requires a new approach. With this in mind, we develop so-called benchmarking tasks to assess ease of learning and ease of use among smartphone platforms. In doing so, we follow the directions taken to develop similar ease of use benchmarking procedures for text editors (Roberts & Moran, 1983; Whiteside, Archer, & Wixon, 1982) and graphic user interface (GUI) operating systems (Gaylin, 1986).

Thus, we identify benchmarking tasks that can be used to evaluate the ease of use and ease of learning of various mobile platforms. In doing so, we recognize that smartphone devices may be used by both novices (i.e., those who have never owned a smartphone) and experienced experts. Because needs over time will evolve with expertise, we develop and apply unique benchmarking task sets to each of these two groups of potential adopters.

To develop these benchmarks, we conducted a Delphi study in three stages (Delbecq, Van de Ven, & Gustafson, 1975; Turoff & Hiltz, 1996) through which we identified and evaluated candidate tasks. As a result of this this process, we developed three task lists, one applicable to smartphone non-users (18 tasks), one applicable to smartphone users (16 tasks), and one general list generated using data from all respondents (16 tasks). As expected, our results show a strong correlation between tasks considered important in determining smartphone ease of use and tasks for which smartphones are frequently used. Our results also show that, although there are seemingly countless possible tasks for which smartphones may be used, certain tasks are carried out with a significantly higher frequency than others.

We then demonstrated the use of these task lists by conducting an exploratory test of the four major, current platforms with a sample of the benchmark tasks: we employed volunteer novice users to assess ease of learning and simulated experts to assess ease-of-use. We simulated the performance of expert users via the keystroke model (Card, Moran, & Newell, 1980, 1983), which Gray, John, and Atwood (1992) found to be highly correlated with actual, experienced users of a particular technology. This finding saves valuable research time in developing and assessing an interface. In our simulation, we used keystroke parameters developed for specific application to mobile devices by Holleis, Otto, Hussmann, and Schmidt (2007) and Holleis, Scherr, & Broll (2011).

This study's results give researchers and practitioners alike guidance on how to evaluate the ease of use and ease of learning of mobile platforms such as smartphones. We also provide a rare examination into how the two evaluations differ. Finally, we contribute to the literature and to practice by demonstrating a

Delphi-based approach to identifying appropriate tasks to be included in benchmarking procedures for future technologies.

2. Background

Computing device interface usability benchmarking has long been a subject of study among information systems researchers. The tradition originated with studies from Roberts and Moran (1983) and Whiteside et al. (1982), who focus on the types of effort required for individuals using text editors and the individual differences that affected their ability to learn a new interface. Gaylin (1986) later extended this research by applying a similar approach to understanding usability through a benchmarking study of a GUI-based personal computer system in which the author evaluated the system based on user performance across several tasks.

In what has become a major touchstone for usability and interface researchers, Card, Moran, and Newell (1980) developed a “keystroke-level model” to evaluate system interaction speed, and tested usability of text editing software by counting error-free keystrokes. They later expanded this work into what they term the goals, operators, methods, and selection rules (GOMS) method for evaluating the human-computer interface (Card, Moran, & Newell, 1983). This method recommends five dependent variables for assessing usability: time, accuracy, learning, functionality, and recall. It views user interactions as consisting of component parts such as physical actions, cognitive processes, and so on. For a user interacting with a spreadsheet, for instance, an operation may consist of identifying the cell into which to enter information, clicking on the appropriate cell, entering the keystrokes that comprise the required information, and so on. In this manner, one can evaluate and measure component actions and processes. Later research has further validated the GOMS model (Gray et al., 1992; John & Kieras, 1996).

Whiteside et al. (1982) has also performed similar research: they not only counted keystrokes, but also accounted for the time between keystrokes when using text editors. They research found that the keystrokes themselves accounted for only about half of the time required to complete tasks in that environment. This finding illuminates areas for potential improvement in usability design.

Continuing to focus on text editing, Roberts and Moran introduced a methodology to evaluate text editor usability both in terms of time and errors (Roberts, 1980; Roberts & Moran, 1983). They argue that emphasizing the features and mechanisms of given text editors cannot be the best approach to evaluating relative usability since, ultimately, no two text editors were comparable based on these facets. They asserted instead that time and errors measured against a common set of tasks yields a clear understanding of a text editor’s usability levels relative to that of an alternative editor. In other words, they recommend evaluating a text editor’s ease of use by measuring its user’s ability to complete objective tasks and not on features or tools that may be idiosyncratic among editors.

This approach, however, required the authors to identify tasks appropriate for text editor comparison. In their study, Roberts and Moran (in Roberts, 1980, and Roberts & Moran, 1983) therefore created a taxonomy of potential tasks: they first identified 212 candidate tasks that could be accomplished through available text editors, then narrowed this set down to 53 critical tasks to be used for actual benchmarking. They then used this task set to evaluate text editors: they asked subjects to complete the tasks and measured the time it took for them to complete them and the number of errors they made. Using this method, they found significant differences in usability among text editors both in terms of time and accuracy. These results were later confirmed in a replication study (Borenstein, 1985).

Later, Gaylin (1986) implemented a similar methodology to evaluate the usability of GUI-based operating systems. In this case, he identified candidate tasks by observing computer operators and noting the frequency with which given functions and commands were used. Following this, he evaluated tasks through interviews and surveys administered to the computer operators, whom he asked to evaluate the candidate commands based on frequency of use, usefulness, friendliness, complexity, naturalness, and importance. He used these data to evaluate the extent to which frequency of use and ease of use were correlated, and analyzed the results from the observations, interviews, and data to create a set of benchmarking tasks.

While the above-mentioned studies provide useful sets of benchmarking tasks, there are key differences between mobile device operating systems in our current study and GUI-based PC operating systems and text editors in previous studies. Smartphones rely on a set of inputs that differ from previous devices. While early PCs required users to input text-based commands and more-modern PCs used (and still use) mouse clicks in a desktop and GUI paradigm, contemporary touch-screen mobile devices are manipulated via tapping, pinching, swiping, and even gyroscopically-determined device positioning. Further, while PCs have traditionally been used for productivity-focused activities such as creating text documents, calculating with spreadsheets, and managing databases, smartphones are more often used for communications and entertainment. With such critical differences, then, assessing smartphone usability cannot depend on tasks identified for PC use. Thus, to apply a similar methodology to determining smartphone ease of use and adoption requires one to identify a new set of benchmarking tasks specifically tailored to the platform.

One final concept is the differentiation between ease of learning and ease of use. According to Mayhew (2013), ease of learning measures how intuitive an interface is for new users, while ease of use indicates how quickly memorized operations can be executed in accomplishing a task. Mayhew raises the issue of interfaces that must be used by both experts and novices, and notes the long-standing use of keyboard shortcuts (for experts) along with pull-down menu choices (for novices) as a method of serving both sets of users.

The smartphone market's newness is its most striking attribute, and the use of the word "market" emphasizes a push by proponents of each platform to sell phones to people who are not necessarily familiar with that platform. Therefore, if the feature set focuses more on novices than experts, it will presumably be easier to attract those users to that platform. Therefore, we also examine the consistency between ease of learning and ease of use.

3. Methodology—Benchmarking Tasks

3.1. Approach

In creating a new set of tasks, we followed other authors' (Gaylin, 1986; Roberts, 1980; Roberts & Moran, 1983) general approach. First, we identified candidate tasks, then narrowed this task set by determining the relative importance of each in contributing to users' perceptions of ease of use.

We were interested in analyzing ease of use from expert users' perspective and ease of learning from novice users' perspective. Given the already widespread adoption of smartphones and seemingly endless stream of announcements of new models on each platform, many of today's smartphone buyers are in fact switching from one platform to another (e.g., Siegal, 2014). Even those who are not changing operating systems in doing so (e.g., moving from one Android-based smartphone to another) may nevertheless be switching to a new brand, new model, new screen size, new operating system version, and so on. In this case, an individual's perceived ease of use of a new phone may be salient to that individual making a selection.

On the other hand, there remains a significant number of consumers who do not currently use smartphones and, as such, may be new to a platform. Such individuals may be unfamiliar with new functions available through these platforms or may place a different emphasis on tasks with which they are more familiar.

We identify an appropriate set of tasks for each of these two groups (and a set that could be applied to both groups) by using the Delphi method (Delbecq et al., 1975; Turoff & Hiltz, 1996). Under this method, panels of experts are consulted to obtain opinions and refine responses to obtain a consensus through intermediated communications. Such studies typically consist of multiple rounds. In the first round, participants brainstorm ideas; in subsequent rounds, the existing idea set is refined and additional ideas may be added. Under some implementations, this process is continued until a steady state of refined ideas is reached. Such processes can be administered in person or, as in this study, through a computer-mediated environment (Turoff & Hiltz, 1996). This method has been used in numerous studies in the information systems literature that require expert-generated sets of ideas (e.g., Brancheau, Janz, & Wetherbe, 1996; Dickson, Leitheiser, Wetherbe, & Nechis, 1984; Holsapple & Joshi, 2000; Schmidt, Lyytinen, Keil, & Cule, 2001).

3.2. Sample

In determining expertise in performing smartphone tasks and their bearing on ease of use, we considered those both experienced and inexperienced with smartphones to be efficacious at perceiving ease of use and ease of learning. We recruited participants in the study via email and social networking via Facebook. We encouraged first-order recruits (i.e., those who we contacted directly) to recruit additional participants. We incentivized participants with a reward drawing that offered them entry into a drawing to win one of five $100 cash cards. We required participants to complete all data-collection phases to qualify for the drawing.

We collected data in three phases (initial identification, refinement, and further refinement). We required that participants complete surveys in all three phases to qualify for entry into the drawing. The first phase was available for 14 days, while the second and third phases were made available to participants for 10 days each.

The initial phase of data collection drew 188 participants. From this initial group, 152 completed the survey for the second phase, and of those 152, 146 completed the final, third-phase survey. Among this final number, 53.3 percent were male, 32.8 percent were students, and 83.9 percent identified themselves as current smartphone users. Among current smartphone users who identified their smartphone's operating system, 44.9 percent used an Apple iOS device, 38.1 percent Android, 11.0 percent Blackberry, and 6.0 percent another operating system.

3.3. Data Collection

3.3.1. Phase One

In the first survey, we asked participants to provide demographic information (age, gender, income level, education level, place of residence) and information regarding their use of information technology (e.g., PC operating system used, mobile operating system used if any, etc.). In addition, we asked participants to suggest up to ten tasks that they considered important in evaluating a smartphone's ease of use. This produced 989 total suggestions (for a mean of 5.46 suggestions per participant); we included suggestions from all participants regardless of whether they continued into successive phases.

A wide audience of potential testers needed to understand and address the tasks included in our final task set. With this in mind, we evaluated the full set of tasks to complete the first-phase list of candidate tasks. Given the need to create the broadest list possible, we included all suggestions, regardless of their frequency of appearance. Some suggested tasks were very common among participants; we found variations of "place phone call", for instance, in 94 of the 188 submissions. Others, however, were unique (e.g., "hdmi output", "using an enhanced reality app").

Some common issues arose among the suggested tasks that required additional judgment. Some entries constituted "composite tasks" that required two or more discrete tasks be carried out; we considered the suggestion "text message", for instance, lacking a verb, as a composite task, which includes both reading and composing text messages. Following Gaylin (1986), who included only discrete and individual tasks in his study's task set, we split such composite tasks into their component parts and included them as two or more tasks. We therefore counted the "text message" task both as "read text message" and "enter/send text message". Other similar composite tasks were decomposed into individual tasks.

Further, some suggestions gave qualitative or overly specific requirements. In these cases (e.g., "make calls promptly"), we ignored the qualitative modifying component; thus, we registered "make calls promptly" the same as "make calls". Finally, some suggestions were too ambiguous to be readily understood or interpreted (e.g., "using the tools") and were omitted from the phase one candidate task list.

At phase one's conclusion, our panel of users and potential users had a list of 106 discrete, independent, actionable tasks. We then paid this task set further consideration in phase two.

3.3.2. Phase Two

We invited the participants who completed the phase one survey to participate in the phase two survey (i.e., we did not allow anyone who did not participate in phase one to participate in phase two). In this second phase, we presented participants with the list of 106 candidate tasks and asked to rate how important each task was for evaluating the ease of use of a smartphone using a seven-point Likert scale ranging from “no importance” to “extremely high importance”. Respondents could also answer with “I am not familiar with this” for each of the candidate tasks.

At the end of the phase two survey, we invited participants to suggest additional tasks that they felt were important to evaluating smartphones’ ease of use and were not generated in phase one. Further, we asked them to suggest ways to make any of the phase one candidate tasks more specific if necessary. As a result, participants suggested an additional 11 tasks (e.g., “manage app-specific security settings”, “set up parental control”). We then added these tasks to the candidate list, which raised the total number of candidate tasks to 117.

3.3.3. Phase Three

We then asked phase two participants to participate in phase three; again, we only allowed those participants who participated in phase two to participate in phase three. In this phase, participants rated how important each 11 newly added task was for evaluating smartphone ease of use, again using seven-point Likert scales. In addition, we asked participants to rate each of the 117 candidate tasks based on how often they performed each. These responses used six-point scales that ranged from “have never” to “hourly”.

4. Analysis—Benchmarking Tasks

4.1. Task Lists

At this point, we used the results from phase two and phase three to create three sets of tasks recommended for evaluating the ease of use of smartphone devices: one set for all users, one for smartphone users, and one for smartphone non-users. Note that two subjects did not self-identify as either users or non-users; we thus included their responses only in the all users list. We generated these lists by calculating the mean importance rating for each task within each target group. These final lists include only those items whose means were one standard deviation or more above their respective means (see Tables 1, 2, and 3).

Table 1. All Users Task List (n = 146, μ = 4.46, σ = 0.79)

	n	Min.	Max.	Mean	S.D.
Check/read email	141	1	7	6.18	1.23
Send email	144	1	7	6.02	1.21
Answer phone call	146	3	7	5.76	1.23
Look up a contact	143	2	7	5.75	1.2
Call a contact	145	2	7	5.68	1.22
Add a contact	143	2	7	5.67	1.17
Conduct an Internet search	145	1	7	5.67	1.39
Send a text	145	1	7	5.65	1.61
Read a text	146	1	7	5.57	1.63
Browse the Internet	145	1	7	5.53	1.4
Take a photo	146	1	7	5.5	1.42
Connect to wireless	144	1	7	5.49	1.51
Get driving directions	142	1	7	5.48	1.45
Check calendar	145	1	7	5.43	1.64
Listen to voice mail	145	1	7	5.3	1.29
Edit a contact	143	2	7	5.28	1.26
Add an appointment	144	1	7	5.27	1.57
Hang up phone call	145	1	7	5.27	1.45

Table 2. Smartphone Users Task List (n = 119, μ = 4.58, σ = 0.80)					
	n	Min.	Max.	Mean	S.D.
Check email	115	3	7	6.37	0.94
Send email	118	3	7	6.16	1.03
Internet search	118	1	7	5.86	1.2
Look up a contact	117	2	7	5.86	1.12
Answer call	119	3	7	5.82	1.17
Call a contact	118	2	7	5.77	1.15
Add a contact	117	2	7	5.77	1.14
Send text	118	1	7	5.73	1.58
Browse Internet	118	1	7	5.7	1.22
Connect to wireless	117	1	7	5.69	1.36
Take a photo	119	1	7	5.69	1.3
Check calendar	118	1	7	5.69	1.48
Get driving directions	116	1	7	5.65	1.36
Read text	119	1	7	5.64	1.61
Add an appointment	118	1	7	5.51	1.45
Read a document	118	1	7	5.42	1.23

Table 3. Smartphone Non-Users Task List (n = 25, μ = 3.88, σ = 0.79)					
	n	Min.	Max.	Mean	S.D.
Send email	24	1	7	5.29	1.73
Check/read email	24	1	7	5.25	1.92
Activate speaker phone	25	2	7	5.24	1.42
Call a contact	25	2	7	5.24	1.45
Answer call	25	3	7	5.24	1.42
Listen to voice mail	25	4	7	5.2	1.26
Add a contact	24	3	7	5.17	1.24
Send text	25	1	7	5.16	1.70
Look up a contact	24	2	7	5.12	1.42
Read text	25	1	7	5.12	1.67
Turn on/off	25	1	7	5.08	1.71
Edit contact	24	3	7	4.92	1.21
Switch between apps	25	1	7	4.84	1.65
Call using hands-free	25	2	7	4.8	1.50
Manage contacts	24	2	7	4.75	1.42
Lock/unlock phone	22	2	7	4.73	1.55
Change ringer volume	24	1	7	4.71	1.68
Change speaker volume	25	2	7	4.68	1.49

Participants self-identifying as smartphone users identified 16 tasks using the same one-standard deviation inclusion criterion, while non-users identified 18. The list inclusive of responses from all participants, again using the same criterion, includes 16 tasks. The smartphone user task list and non-user task list include eight items in common: check email, send email, look up contact, answer call, call a contact, add a contact, send text, and read text.

As expected, the list generated from smartphone non-users' responses included seven phone-specific tasks that were not included in the smartphone user list: activate speaker phone, listen to voice mail, edit a contact, hands-free calling, manage contacts, lock/unlock phone, and change ringer volume. Interestingly, this group also identified two tasks that could be considered smartphone-specific functions: switch between apps and lock/unlock phone.

The list generated from smartphone users' responses includes eight tasks absent from the non-user list: search the Internet, browse the Internet, make a wireless connection, take a photo, check the calendar, get driving directions, add an appointment, and read a document. Again, as expected, these tasks could be construed as very smartphone specific. It may be that smartphone non-users do not have enough experience with the technology to appreciate the need for such tasks.

4.2. Task Importance vs. Task Frequency

In addition, we wanted to understand the possible influence between the frequency of task performance and participants' perceptions of task importance in evaluating smartphone ease of use. Thus, we compared the means of the ease of use importance scores among the 117 candidate tasks with the frequency of use. We found a significant correlation between these metrics (*Pearson's* $r = 0.823$, $p < 0.001$). Figure 1 visually corroborates this finding with a scatterplot.

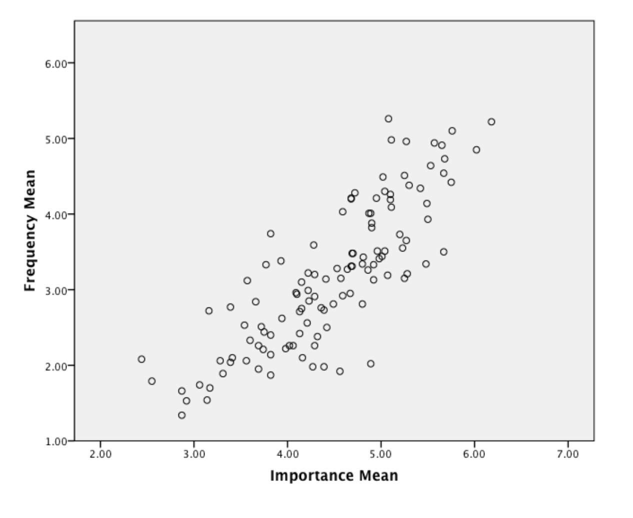

Figure 1. Importance Mean vs. Frequency Mean

However, given that our measurement of task performance frequency was not a ratio scale, we note that this approach may be insufficient to establish significant correlation. Thus, we converted ease of use importance values and frequency of use ratings into rankings by sorting the 117 tasks based on each of these values. Using these rankings, we produced an additional scatterplot graph (see Figure 2). Using Kendall's tau, we again found a significant correlation between frequency of use and perception of ease of use importance ($\tau = 0.651$, $p < 0.001$).

4.3. Task Frequency

In addition to this relationship, we also sought to understand participants' task use frequency in isolation. To this end, we first compiled the tasks most frequently used by our sample of smartphone users; Table 4 shows tasks with use frequency means that were more than one standard deviation above the grand mean. In addition, Figure 3 shows the distribution of means.

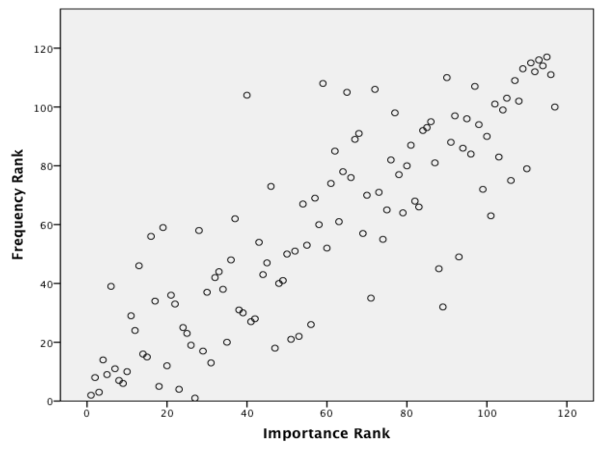

Figure 2. Frequency Rank vs. Importance Rank Scatterplot

Table 4. Frequency of Task Performance Among Smartphone Users

	n	Min.	Max.	Mean	S.D.
Check/read email	112	1	6	5.49	.684
Check time/date	112	2	6	5.41	.833
Lock phone	114	1	6	5.25	1.261
Answer call	115	3	6	5.17	.634
Send email	113	1	6	5.12	.753
Read text	115	1	6	5.11	.915
Send text	113	1	6	5.07	.961
Hang up phone call	115	1	6	5.05	.782
Call a contact	114	3	6	4.88	.718
Browse Internet	113	1	6	4.82	.956
Open an app	112	1	6	4.79	1.297
Internet search	113	1	6	4.70	.972
Check calendar	111	1	6	4.66	1.124
Check weather	113	1	6	4.65	.810
Find a contact	114	2	6	4.61	.816
Switch between apps	110	1	6	4.45	1.385
Check call history	115	1	6	4.43	.992
Adjust speaker volume	114	1	6	4.43	1.212
Listen to voice mail	115	2	6	4.41	.826
Check news	113	1	6	4.38	1.190
Connect to wireless	112	1	6	4.38	1.409
Call by dialing number	115	2	6	4.36	.870
Ignore incoming call	115	1	6	4.35	1.060
Use speaker phone	113	1	6	4.32	1.120
Adjust ringer volume	114	2	6	4.31	1.191

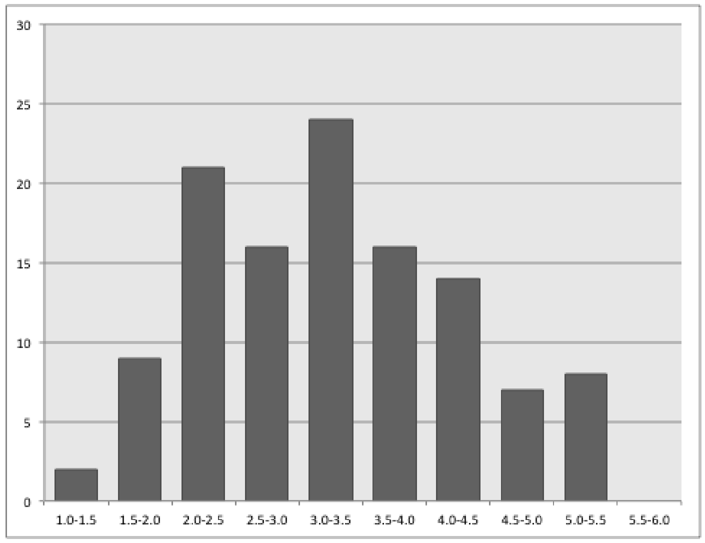

Figure 3. Distribution of Task Performance Frequency Means

In examining these data, we found that a significant difference in frequency of use exists among tasks in this list. In other words, some tasks were reported to be performed more frequently than others. In fact, the task rated the most frequently performed, check/read email, had significantly higher frequency than the fourth most frequently performed task, answer call (t = 4.318, p < 0.001), and all other tasks with lower performance frequency mean values. We found a similar result for the second most frequently performed task, check time/date. This task was performed with a higher frequency than the fourth most frequently performed task, answer call (t = 2.857, p = 0.005), and all other tasks with lower performance frequency mean values.

5. Preliminary Assessment of the Interfaces

5.1. Context

With the benchmark task list established, we then demonstrated the usefulness of the list. Here, we assessed the smartphone interfaces and explored potential differences between those who are experienced on a given platform (“platform experts”) and those who have never used the given platform before (“platform novices”). This second group is particularly interesting in that it likely includes smartphone users who may be evaluating switching to a new smartphone platform. Therefore, the “all users” list in Table 1 will enable us to compare novices and smartphone platform experts on the same list of tasks.

These two groups (platform experts and novices) are somewhat different from smartphone users and non-users. In our preliminary assessment, we were more interested in those who were highly skilled in the given platform, rather than just those who were “users” of that platform. Further, according to Card et al. (1983), the keystroke model assumes expert, error-free performance.

The novice or non-smartphone user group represented a difficult sampling problem. Given the abrupt rampup of smartphone usage saturation, when focusing on applying the benchmarks, it became increasingly difficult to find those who were not experienced with at least one platform of smartphone. Most we were able to find were either uninterested in or unable (due to age or disability) to use smartphones. Those individuals would form a sample that would provide difficulty when trying to generalize our results. Because of these issues, and reflecting the changing market, we chose to categorize users of other platforms as novices in each focal platform.

We performed separate evaluations for each user group. To evaluate the interfaces for platform novices, we asked at least two people who were unfamiliar with each platform to perform each of the benchmark tasks. We recorded these participants on video to allow us to collect timing data. To evaluate the interfaces for platform experts, we evaluated each platform using parameters from the keystroke model (Card et al., 1983) as updated by Holleis et al. (2007, 2011) specifically for the smartphone platform. We applied the parameters to each step of the given tasks.

As such, our time measurements were objective and, as such, provided significant contrast to the subjective evaluations widely available in reviews found in popular media and on the Internet. These subjective evaluations can be problematic because they tend not to consider bias based on reviewers' personal investment and deep experience with a given platform. In addition, these reviewers cannot reasonably have equivalent experience with all platforms. Using these objective timings, we consider performance to be better when tasks are completed in a shorter period of time and worse when completed in a longer timeframe (Card et al., 1983).

Additionally, we note that neither platform novice nor platform expert timings tell the entire story about a smartphone platform's ease of use. Both should be examined. Platform novice performance is important to consider in that it highlights how easy one learns a new platform and may indicate that users are becoming lost or clicking unnecessarily while attempting to perform a task. Such actions may lead to frustration, and, thus, to the novice user deciding to choose a different platform.

Platform expert performance reflects speed of operation. Such performance features two elements: "muscle memory", which has been studied for hundreds of years (Adams, 1987), and error-free performance (Card et al., 1983), hallmarks of extensive experience. Expert performance is important because it indicates the performance that can be expected as a result of using a platform in the long term.

Therefore, we separately analyzed each user group.

5.2. Platform Novice Timings

In finding an available sample of participants to illustrate our set of benchmark tasks, we identified users who were unfamiliar with each smartphone platform to be examined (Android, BlackBerry, iOS, and Windows 8). To collect data, we used a representative device for each of these operating systems: an Android using the "Jellybean" or "Kitkat" version of the operating system, a Blackberry Torch[1], an iPhone 5, and a Windows 8 mobile phone[2].

This illustration of applying the benchmark tasks is not meant to comprehensively or permanently compare the platforms. The sample sizes in our preliminary analyses were small and data from more individuals should be collected to obtain more definitive results. The sample size for each of the platforms above was 6, 6, 4, and 2, respectively. Note, however, that pilot tests over two years using various subsets of the benchmark task list revealed very similar results for the top and bottom performers.

Table 5 provides the results of our preliminary analysis with average timings for each of the benchmark tasks for each platform. The cell containing each platform's best performance is filled in with green and the cell containing each platform's worst performance is filled in with pink.

The best performance overall was found on the iPhone 5, with 316.8 seconds to accomplish all tasks, while the worst performance overall was found on the Blackberry, with 489.6 seconds to accomplish all 19 tasks. The iPhone 5 placed first in 10 of the 19 tasks, while the Blackberry placed last on 9 of the 19 tasks. Interestingly, the Windows 8 phone ranked second in overall time but also had 8 first-place finishes, and the Android phone ranked third, with two first-place finishes. The Blackberry, which finished last, also had the smallest number of first-place finishes (one). For last-place finishes, the iPhone had the smallest number (1), the Android and Windows Phone each had 4, and the Blackberry had 9.

[1] The Blackberry Torch was furnished through the generosity of Verizon Wireless, Cranberry, PA.
[2] The Windows 8 mobile phone (Nokia 620) was furnished through the generosity of the Microsoft Corporation.

Table 5. Platform Novice Timings				
	iPhone	Android	Windows	Blackberry
Check/read email	8.2	13.6	11.5	20.4
Send email	21.4	37.1	16.0	34.3
Answer phone call	1.9	5.4	1.0	2.5
Look up a contact	18.6	12.2	25.5	32.1
Call a contact	10.0	14.6	7.5	15.3
Add a contact	17.9	20.9	42.0	39.9
Internet search	17.5	14.4	11.0	37.5
Send a text	12.7	18.3	16.5	29.9
Read a text	3.6	6.8	7.5	9.7
Browse the Internet	23.9	18.0	12.0	43.0
Take a photo	11.2	14.1	40.0	16.1
Connect to wireless	24.1	17.3	13.0	30.0
Get driving directions	43.4	60.2	32.0	56.3
Check calendar	16.9	11.8	29.0	8.4
Listen to voice mail	31.6	42.9	8.0	64.4
Edit a contact	17.7	38.2	35.0	27.8
Add an appointment	34.8	14.9	24.5	31.3
Hang Up phone call	1.3	3.1	3.5	3.0
Total	316.8	363.7	335.5	486.7
Rank	1	3	2	4
Number of first places	8	2	8	1
Number of last places	1	4	4	9

5.3. Platform Expert Timings

We calculated expert timings for completing tasks using the keystroke model (Card et al., 1980, 1983). Under the keystroke model, each action an expert user takes is assigned a time value based on rigorous observations. The time it takes to complete a given task, then, is calculated as equivalent to the sum of the time values for the required actions. We evaluated expert timings from two separate studies by Holleis et al. (2007; 2011) using a set of original and modified parameters, respectively, developed specifically for smartphone users (see Table 6). We used only a subset of these operators, including only those operators needed to complete the tasks tested.

Table 6. Selected Keystroke-Level Parameters (from Holleis et al., 2007, 2011)		
Abbreviation	Action	Time(seconds)
F – Finger movement	Moving a finger on the device, even from one keyboard character to another	.23 s
K – Keystroke	Press a key (average time)	.39 s
G – Gesture	Dragging	.80 s
M – Mental	Mental preparation for a subtask	1.35 s
P - Pointing	Moving the mobile device to the proper orientation	1.00 s
S_{Micro} – Micro attention shift	Moving eyes from the keyboard to the display to check input	.14 s

Holleis et al. (2011) also provides a set of rules for placing mental operators. In using these timings, Holleis et al. specifies an important rule regarding mental preparation (M) that significantly affects our timings. A mental preparation (M) must occur immediately prior to a keystroke (K) or gesture (G), unless there are multiple keystrokes in succession, in which case the M occurs only before the first of the set of keystrokes. Schulz (2008) illustrates steps in which M times are inserted into a set of steps obtained by observing the interactions of an expert.

We consulted an expert user of each platform to compile a comprehensive set of low-level steps necessary for each task. We then expanded each list of steps to include the Holleis et al. (2011) F, M, and S_{Micro} steps and all timings listed in Table 6. Table 7 illustrates how we elaborated on one task on the iPhone in detail[3].

Table 8 provides the total time measurements we compiled using the operators determined for each task (see Appendix), multiplied by the parameters in Table 6. A very short text and email message (10 characters) was used, in which subject and location content each contained seven characters, a 10-digit phone number was employed for all calls and texts, and a frequently-used email address identifiable after typing two characters was used to standardize the data entered. We did not attach an M timing to the use of the Blackberry "black key[4]"; platform-expert users would be accustomed to pressing the key automatically (without thinking and thus without accruing any mental preparation time).

Table 7. Illustration of Adding a Contact with the iPhone

Sub-task	Operator	Time, each	Quantity	Time
Prepare for task	m	1.35	1	1.35
Click contact icon	k	0.39	1	0.39
Prepare for next task	m	1.35	1	1.35
Click + sign upper corner	k	0.39	1	0.39
Prepare for next task	m	1.35	1	1.35
Click first name field	k	0.39	1	0.39
Type first name (5 characters)	f	0.23	5	1.15
Check screen	S_{Micro}	0.14	1	0.14
Prepare for next task	m	1.35	1	1.35
Click last name icon	k	0.39	1	0.39
Type last name (5 characters)	f	0.23	5	1.15
Check screen	S_{Micro}	0.14	1	0.14
Prepare for next task	m	1.35	1	1.35
Click "phone"	k	0.39	1	0.39
Type phone number (10 digits)	f	0.23	10	2.3
Check screen	S_{Micro}	0.14	1	0.14
Prepare for next task	m	1.35	1	1.35
Click DONE upper right corner	k	0.39	1	0.39
Total Time				**15.46**

[3] Complete set of spreadsheets for all four platforms are available on request from either author.
[4] The "black key" is a prominent, unlabeled key that serves as "enter" to complete an operation.

Table 8. Keystroke-Level Model Timings Representing Experts				
	iPhone	Android	Windows	Blackberry
Check/read email	3.62	3.62	3.62	2.11
Send email	16.12	16.12	16.12	12.13
Answer phone call	1.74	2.15	1.74	1.74
Look up a contact	8.97	5.68	3.94	3.12
Call a contact	9.11	9.30	3.62	3.35
Add a contact	15.46	13.58	19.33	14.92
Internet search	8.32	5.23	6.97	4.02
Send a text	9.40	9.40	11.74	5.49
Read a text	2.35	2.35	3.23	1.20
Browse the Internet	10.86	9.12	10.72	7.84
Take a photo	5.83	5.83	7.44	5.90
Connect to wireless	13.55	16.09	13.41	11.50
Get driving directions	9.40	7.66	9.65	13.15
Check calendar	1.74	1.74	1.74	.92
Listen to voice mail	1.74	5.50	2.13	5.16
Edit a contact	12.78	13.93	13.33	8.05
Add an appointment	19.28	22.36	21.58	12.32
Hang up phone call	1.74	1.74	1.74	1.74
Total	155.11	151.40	150.70	114.66
Rank	4	3	2	1
Number of first places	3	3	1	13
Number of last places	6	9	7	1

In assessing ease of use for platform experts, Table 8 notes the Blackberry as the fastest platform. The Blackberry took the least total time to complete all tasks and garnered 13 first-place finishes with only one last-place finish. In contrast, the iPhone took the longest among the four platforms.

Comparing Tables 5 and 8 illustrates the differences that are found for platform novices and platform experts in comparing devices. The Blackberry depends less on searching for needed icons and more on memorized sequences for the most commonly used functions. The sequences can be launched nearly automatically by the expert user on most occasions. In contrast, for platform novices, the iPhone was quicker for finding the proper icons and functions needed to perform the benchmark tasks, while the Blackberry took longest for novices to perform those tasks.

6. Discussion and Limitations

In this paper, we established and illustrate the use of a set of benchmarking tasks for assessing smartphone ease of use and ease of learning. Using these task sets, future investigators can evaluate smartphone platforms using objective criteria (e.g., task completion time, as illustrated in this paper, and accuracy). Researchers should not limit their scrutiny to subjective criteria (e.g., idiosyncratic perceptions of ease of use and ease of learning) despite these being commonly used in comparisons found in blogs, social media, magazines, and newspaper columns.

Further, these benchmarks can be of significant value for practice. Tests based on these task sets could form the basis for evaluating new operating system enhancements from users' perspective. Further, creators of new mobile applications could use findings facilitated by these task sets in developing new software and new approaches to usability problems.

Our demonstration shows that these task sets can be used effectively. While limited in scope and predictive power, we found that smartphone ease of learning and ease of use should be considered separately because they yielded different results, implying different best practices. Ease of learning should be assessed using platform novices who have no experience with the platform under scrutiny. Ease of use, on the other hand, should be assessed by either measuring platform experts through a methodology similar to the keystroke model used here or through direct laboratory data collection.

Interestingly, in our results, the four most well-known platforms yield starkly contrasting outcomes when used by platform novices as opposed to platform experts. Based on our keystroke model findings, Blackberry is fastest among all platforms for those very familiar with that platform. In contrast, based on our experimental results, the iPhone is fastest for platform neophytes. The Windows platform ranked second and the Android platform ranked third for both groups.

That said, we note a specific limitation to our study; namely, that the steps required for the benchmark tasks for the keystroke model timings were specified by a relatively small sample of available actual users. A different sample of users specifying these required steps could specify them differently and, thus, yield different outcomes based on the keystroke model timings. Because of this limitation, the results of our demonstration should be considered tentative, and replication with more subjects and further detailed study are needed to yield definitive results.

Also, there are some specialized applications that are not considered here. Some organizations might deploy smartphones for a specific, limited set of tasks. For example, picking inventory, delivering goods, or recording medical procedures might be the most organizationally relevant tasks in a particular situation. In those cases, that organization might develop its own list of crucial tasks and/or subtasks to be performed by employees. They might also prefer to take a more long-term perspective and thus be more concerned with ease of use rather than ease of learning. In that case, a customized list of benchmark tasks should be used along with keystroke model timings in making a platform-adoption decision. Following adoption of a platform, suitable time should be allowed to adequately trail individuals to reach a high level of expertise as soon as possible.

Although this study was based on relatively small samples, we are confident that these results sufficiently demonstrate the value of benchmark tasks, the use of keystroke model timings published previously in the literature, and the differentiation of ease of learning and ease of use when considering smartphone usability.

When choosing a smartphone platform, it is important to assess whether timings are important over the long term, short term, or both, and, thus, the relative importance of ease of use versus ease of learning. Users who invest more time in memorizing key sequences might be able to use a smartphone more efficiently. Those who need to examine sets of icons for every operation might be able to startup faster, but the potential for more numerous screens, slower speed, and/or lack of shortcuts might make tasks take longer as they use the device over a longer period of time. Practitioners and other researchers will be able to make use of our approach to arrive at further insights that expand our understanding of smartphone usability.

References

Adams, A. J. (1987). Historical review and appraisal of research on the learning, retention, and transfer of human motor skills. *Psychological Bulletin, 101*(1), 41-74.

Borenstein, N. S. (1985). The evaluation of text editors: A critical review of the Roberts and Morgan methodology based on new experiments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vol. 16, pp. 99–105). New York, NY: ACM.

Brancheau, J. C., Janz, B. D., & Wetherbe, J. C. (1996). Key issues in information systems management: 1994-95 SIM Delphi results. *MIS Quarterly*, *20*(2), 225–242.

Card, S. K., Moran, T. P., & Newell, A. (1980). The keystroke-level model for user performance time with interactive systems. *Communications of the ACM*, *23*(7), 396–410.

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Columbus, L. (2013). IDC: 87% of connected devices sales by 2017 will be tablets and smartphones. *Forbes.com*. Retrieved March 22, 2014, from http://www.forbes.com/sites/louiscolumbus/2013/09/12/idc-87-of-connected-devices-by-2017-will-be-tablets-and-smartphones/

Delbecq, A. L., Van de Ven, A. H., & Gustafson, D. H. (1975). *Group techniques for program planning: A guide to nominal group and Delphi processes*. Glenview, Ill: Foresman Scott.

Dickson, G. W., Leitheiser, R. L., Wetherbe, J. C., & Nechis, M. (1984). Key information systems issues for the 1980s. *MIS Quarterly, 8*(3), 135-159.

Gaylin, K. B. (1986). How are windows used? Some notes on creating an empirically-based windowing benchmark task. *ACM SIGCHI Bulletin*, *17*(4), 96–100.

Gray, W. D., John, B. E., & Atwood, M. E. (1992). The precis of Project Ernestine or an overview of a validation of GOMS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 307–312). New York, NY: ACM.

Holleis, P., Otto, F., Hussmann, H., & Schmidt, A. (2007). Keystroke-level model for advanced mobile phone interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* San Jose, CA: ACM Press.

Holleis, P., Scherr, M., & Broll, G. (2011). A revised mobile KLM for interaction with multiple NFC-tags. In P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, & M. Winckler (Eds.), *Proceedings of the INTERACT Conference, Part IV, LNCS 6949* (pp. 204-221). Lisbon, Portugal: IFIP.

Holsapple, C. W., & Joshi, K. D. (2000). An investigation of factors that influence the management of knowledge in organizations. *The Journal of Strategic Information Systems*, *9*(2-3), 235–261.

John, B. E., & Kieras, D. E. (1996). The GOMS family of user interface analysis techniques: Comparison and contrast. *ACM Transactions on Computer-Human Interaction*, *3*(4), 320–351.

Mayhew, D. J. (2013). Ease of learning vs ease of use—they aren't the same thing. *Bestica.* Retrived September 21, 2014, from http://community.bestica.com/profiles/blogs/ease-of-learning-vs-ease-of-use-they-aren-t-the-same-thing

Nielsen. (2014a). Mobile millennials: Over 85% of generation Y owns smartphones. *Nielsen.com.* Retrieved September 21, 2014, from http://www.nielsen.com/us/en/insights/news/2014/mobile-millennials-over-85-percent-of-generation-y-owns-smartphones.html

Nielsen. (2014b). Smartphone milestone: Half of mobile phone subscribers ages 55+ own smartphones. *Nielsen.com.* Retrieved September 21, 2014, from http://www.nielsen.com/content/corporate/us/en/insights/news/2014/smartphone-milestone-half-of-americans-ages-55-own-smartphones.html

Norman D., & Nielsen J. (2010) Gestural interfaces: A step backward in usability. *Interactions*, *17*(5), 46-49.

Roberts, T. L. (1980). *Evaluation of computer text editors* (Doctoral dissertation). Stanford University.

Roberts, T. L., & Moran, T. P. (1983). The evaluation of text editors: methodology and empirical results. *Communications of the ACM*, *26*(4), 265–283.

Schmidt, R., Lyytinen, K., Keil, M., & Cule, P. (2001). Identifying software project risks: An international Delphi study. *Journal of management information systems*, *17*(4), 5–36.

Schultz, T. W. (2008). *Evaluating mobile phones with the keystroke-level model and other desktop methods* (Master's thesis). University of Oslo. Retrieved March 23, 2014, from https://www.duo.uio.no/handle/10852/9883

Siegal, J. (2014). Android users are deserting in droves ahead of iPhone 6 launch, trade-ins up 200%. *BGR.com.* Retrieved September 22, 2014, from http://bgr.com/2014/09/18/iphone-6-vs-android-users/

Turoff, M., & Hiltz, S. R. (1996). Computer based Delphi processes. In M. Adler & E. Ziglio (Eds.), *Gazing into the oracle: The Delphi method and its application to social policy and public health*. London: Jessica Kingsley Publishers.

Whiteside, J., Archer, N., & Wixon, D. (1982). How do people really use text editors? In *Proceedings of the ACM SIGOA Conference on Office Information Systems* (pp. 29-40).

About The Authors

Dennis Galletta is an AIS Fellow, was previously President of AIS, and is a Professor of MIS at the Katz School, University of Pittsburgh. He also serves as Director of Doctoral Programs at Katz. He obtained his PhD in MIS from the University of Minnesota. His research interests cover end-user behavior, attitudes, and performance, as well as behavioral security. His journal papers have appeared in outlets such as *Management Science*, *MIS Quarterly*, *Information Systems Research*, *Journal of MIS*, *European Journal of Information Systems*, *Communications of the ACM*, *Decision Sciences*, *Data Base*, and *Information and Management*. He has served on several editorial boards, including *ISR*, *MIS Quarterly*, *Journal of MIS*, and *Journal of AIS*, was senior editor of *JAIS*, and is Preeminent Editor at Data Base. He is a founding co-Editor-in-Chief of AIS Transactions on Human-Computer Interaction. His research studies have also been featured in outlets such as Business Week, Wall Street Journal, Computerworld, and CNN Television. He co-published a two-volume collection of research on human-computer interaction in the MIS field for Vladimir Zwass's series on Advances in Management Information Systems. He won a “Developmental Associate Editor Award” at the *MIS Quarterly* in 2006. He launched the development of Special Interest Groups in AIS. He served as the ICIS Treasurer from 1994 to 1998, was a member of AIS Council representing the Americas in 1996 and 1997, chaired the first Americas conference for AIS (and introduced electronic submissions and registration to the IS field), chaired ICIS 2011 in Shanghai, was program co-chair for ICIS 2005 and AMCIS 2003, was Editor-in-Chief of AISWorld from 2004-2006, and was ICIS Doctoral Consortium co-chair in 2008. He has provided keynote speeches at conferences in the US, China, Mexico, Brazil, and on board the Disney Wonder cruise ship. He also taught undergraduate information systems courses on the fall 1999 voyage of Semester at Sea, while sailing around the world.

Brian Kimball Dunn is an Assistant Professor of Management Information Systems at the University of Oklahoma’s Price College of Business. He received his PhD in Information Systems from the Katz Graduate School of Business at the University of Pittsburgh. His research interests focus on user perceptions of digital content environments and platforms. His research has been presented at multiple meetings of both the Hawaii International Conference on System Sciences (HICSS) and the International Conference on Information Systems. He has also co-authored Harvard business cases. Prior to entering academia, he spent ten years in industry having earlier earned a Masters of Business Administration from the University of California, Irvine and a Bachelor of Arts degree in German from Brigham Young University.

ISSN: 1944-3900

Transactions on Human-Computer Interaction