Taylor & Francis
Taylor & Francis Group

An Improved Usability Measure Based on Novice and Expert Performance

Karl F. MacDorman, Timothy J. Whalen, Chin-Chang Ho, and Himalaya Patel
Indiana University School of Informatics, Indianapolis

The novice–expert ratio method (NEM) pinpoints user interface design problems by identifying the steps in a task that have a high ratio of novice to expert completion time. This study tested the construct validity of NEM's ratio measure against common alternatives. Data were collected from 337 participants who separately performed 10 word-completion tasks on a cellular phone interface. The logarithm, ratio, Cohen's d, and Hedges's $ĝ$ measures had similar construct validity, but Hedges's $ĝ$ provided the most accurate measure of effect size. All these measures correlated more strongly with self-reported interface usability and interface knowledge when applied to the number of actions required to complete a task than when applied to task completion time. A weighted average of both measures had the highest correlation. The relatively high correlation between self-reported interface usability and a weighted Hedges's $ĝ$ measure as compared to the correlations found in the literature indicates the usefulness of the weighted Hedges's $ĝ$ measure in identifying usability problems.

1. INTRODUCTION

The usability of a system interface depends on at least three points. According to the International Standards Organization (ISO 9241-11, 1997), "system usability comprises the extent, to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use." Secondary usability concepts include the ease of avoiding errors and understanding and learning the interface. Although user satisfaction is subjective, efficiency and effectiveness can be operationalized using performance metrics (Brinkman, Haakma, & Bouwhuis, 2007; Lindgaard & Dudek, 2003). Efficiency is typically operationalized by such metrics as completion time, time until event, deviation from optimal path, and use frequency, and effectiveness

---

is typically operationalized by such metrics as error rate, binary task completion, spatial accuracy, outcome quality, completeness, and recall (Hornbæk & Law, 2007).

The importance of efficiency depends on the user's goals. For example, a user of a productivity application like a word processor has different goals from a visitor to a game Web site. (For the purposes of this study, a *task* is considered to be achieved by a sequence of steps performed on an interface, and a *step* is performed either by a single correct action or by one or more erroneous actions and undos followed by the correct action.) Efficiency also depends on learnability. A highly learnable interface focuses on the needs of the novice, enabling the novice to gain proficiency and, in turn, efficiency (Nielsen, 1994).

When a productivity interface exploits the existing skills of users, even those who are new to the interface can perform tasks rapidly and with few errors. However, given enough practice, a user can become proficient with an interface that was at first counterintuitive. Hence, a large performance gain resulting from the attainment of expertise can indicate an interface is difficult to use. By the same logic, if novices take much longer than experts to perform a task, this could indicate usability problems with the interface.

The difference between novice and expert performance can indicate problems in a productivity interface, but finding a simple usability measure is difficult. Thorough troubleshooting cannot be achieved by simply comparing the action sequences of a single novice and expert user (e.g., Uehling & Wolf, 1995) or comparing performance measures that are aggregated across all steps in a task (e.g., Dillon & Song, 1997). By comparing novice and expert performance at each step in a task rather than for the whole task, it is easier to identify specific efficiency and effectiveness problems. This approach enables usability engineers to determine accurately where novices encounter trouble and the frequency, number of errors, and duration of the usability problem at that step. Expert performance can be estimated from reliable models of human cognition and movement (e.g., CPM-GOMS and Fitts's Law; John & Kieras, 1996; MacKenzie, 1992; Silfverberg, MacKenzie, & Korhonen, 2000). Software tools are available to automate this process (e.g., CogTool, based on KLM-GOMS; Luo & John, 2005).

The novice–expert ratio method (NEM) compares the average time it takes a group of novice users to complete one step of a task to that of an expert group (Urokohara, Tanaka, Furuta, Honda, & Kurosu, 2000). The expert group provides an estimate of each step's minimum completion time. A large ratio indicates the novice group is taking much longer than the minimum time required. Because the novice-expert (NE) ratio is calculated for each step in a sequence of actions, usability engineers can use the NE ratio as a diagnostic measure to determine which steps demand additional time from novices, enabling the engineers to identify and correct the specific aspects of a user interface that are counterintuitive.

Urokohara et al. (2000) claimed that software engineers and industrial designers in Japan tend to find it easier to accept the NE ratio than measures derived from self-reported information, such as usability questionnaires or heuristic evaluation, because they deem the self-reported information to be more affected by individual differences and personal bias (see also Kuniavsky, 2003). Questionnaires typically require testers to evaluate an interface by stating their level of agreement with

statements concerning various aspects of its usability (Brooke, 1996; Chin, Diehl, & Norman, 1988). For example, Lewis (1995) applied psychometric methods to develop and evaluate the IBM Computer System Usability Questionnaire. Its 19 questions loaded on three factors: *system usefulness, information quality,* and *interface quality*. Each factor had high reliability, with Cronbach's alphas of .93, .91, and .81, respectively. These three factors accounted for 98.6% of rating variability in his study. An analysis of 5 years of usability studies has confirmed these results (Lewis, 2002).

Heuristic evaluation engages a group of evaluators—usually usability engineers—in testing an interface's adherence to recognized usability principles (Desurvire, Lawrence, & Atwood, 1991; Nielsen, 1992). Examples include consistency, feedback, error prevention, and reduction of memory load (Molich & Nielsen, 1990; Nielsen & Molich, 1990). The goal is to ensure that the interface is easy to learn, pleasant to use, and effective in performing the intended tasks (Gould & Lewis, 1985). Heuristic evaluation is useful for diagnosing usability problems and finding solutions to them. The process, however, is open to differences in interpretation (Faulkner & Wick, 2005). Different evaluators can produce varying estimates of the time required to complete a task (Nielsen & Phillips, 1993). Cognitive walk-throughs, think-aloud protocols, and other usability evaluation methods face the same limitations, and it is difficult to compare their effectiveness accurately (Gray & Salzman, 1998). A review of 11 usability studies revealed that average evaluator agreement ranged from 5 to 65% for heuristic evaluation, cognitive walk-throughs, and think-aloud protocols (Hertzum & Jacobsen, 2003).

1.1. Deficiencies in NEM

Although objective performance measures are more expensive to collect than user-reported preferences or the heuristic analyses of usability engineers, the rewards of developing a faster interface are potentially much greater (Nielsen & Phillips, 1993). The NEM provides such an objective performance measure to pinpoint usability issues; however, it has three major shortcomings:

1. User error and recovery rates are important quantitative measures of usability, but they are not included in calculating the NE ratio.
2. It is unclear how to calculate the NE ratio when a user fails to complete a step.
3. The merits of the ratio measure have never been appraised relative to other possible measures, such as standard statistical measures of effect size.

Let us consider these points in turn.

The first limitation of the NE ratio is that it is derived solely from completion time data, overlooking error and recovery rates, which can provide valuable information about an interface's usability. Even if a task can be performed more quickly with one interface than another, users may feel frustrated if the error rate is higher. Frustration has been identified as a major issue for novice users (Bessiere, Newhagen, Robinson, & Shneiderman, 2006; Kang & Yoon, 2008; Kjeldskov, Skov,

& Stage, 2005). In addition, subjective measures of satisfaction are more strongly correlated with error rate than with completion time (Hornbæk & Law, 2007; Sauro & Lewis, 2009). Therefore, it makes sense to consider not only how long it takes a user to finish a step but whether the user made mistakes in planning or slips in execution and had to undo and redo actions (Goonetilleke, Shih, On, & Fritsch, 2001). It is also possible to create a single measure that combines information about completion time and the number of actions required to complete a step. (The heightened correlation of this kind of measure with self-reported usability is shown in Figure 2.)

Second, when a novice user fails to complete a required step, there is no novice completion time value to use in calculating the NE ratio. The developers of NEM have suggested that in this case, a major usability problem is apparent, so it is unnecessary to calculate a particular NE ratio. However, novice users may fail to complete a step for reasons other than the design of the interface (e.g., conflicting cognitive and motivational strategies; Carroll & Rosson, 1987). In a study with a large sample size, the poor performance of a single user should not force the rejection of a design.

Third, concerning the relative merits of the ratio measure, it is worth noting that other measures have more commonly been applied to participant time measurements (e.g., Ratcliff, 1993). These include Cohen's d, Hedges's $\hat{g}$, the logarithm transformation, and the reciprocal transformation.

1.2. Alternatives to the Ratio Measure

A strength of NEM is that its ratio output is inherently independent of the absolute task time (Urokohara et al., 2000). Two statistical measures from the social sciences, Cohen's d and Hedges's $\hat{g}$, can also measure the relative difference between two groups' proficiency with an interface. Cohen's d is a standard measure of effect size in comparing groups. Cohen's d compensates for the correlation in each group's magnitude differences and its underlying variability by dividing the difference in each group's mean by the average of each group's standard deviation (Cohen, 1977). Cohen's d is an accurate measure of effect size if the sample sizes of the two groups or their standard deviations are equal. However, it can overestimate or underestimate the effect size if one group is larger than the other and the standard deviations are different.

Hedges's $\hat{g}$ is similar to Cohen's d in how it measures effect size, with the exception that a correction is added to adjust for dissimilar sample sizes and standard deviations of the groups being compared (Hedges & Olkin, 1985). A novice group tends to be more heterogeneous than an expert group with respect to performance with a particular user interface, because skill transfer from experience with other related interfaces has a larger influence on novice performance (Norman, 1983). Each novice's performance is a reflection of that particular individual's prior experience, whereas the experts' extensive practice decreases variability as they approach their performance limits. Therefore, the standard deviations for the novice and expert completion times and number of actions taken tend to differ, which is a justification for using Hedges's $\hat{g}$. In sum,

Hedges's $\hat{g}$ eliminates the need to assign an equal number of participants to each group, which has been the practice to simplify analysis (Faulkner & Wick, 2005).

The logarithm and reciprocal transformations have been used to improve central tendency estimates by removing positive skew from the data (Ratcliff, 1993). Completion time distributions typically have extended upper tails, because some participants take a long time to complete a task or fail to complete it altogether. Removing the positive skew is helpful when applying parametric statistical tests, because these tests assume a normal (symmetrical) distribution. Although the median is seldom used for completion time data, it can provide a useful baseline measure for gauging the quality of other measures.

1.3. Evaluating Measure Calculation Method Validity

Correlations with other related concepts can be used to compare the construct validity of different measures. (This is not the typical application of construct validation to the assessment of different ways of measuring the same construct; rather it is the less common application of construct validation to the assessment of different ways of calculating a measure of a construct; see Greenwald, Nosek, & Banaji, 2003.) For example, in a sufficiently large sample of people, mass and height are generally positively correlated. Thus, a typical person's mass in kilograms should correlate more highly with his or her height measured to the nearest centimeter rather than to the nearest decimeter. Using the same approach, we can evaluate the construct validity of different ways of calculating novice-expert measures from the same performance data by comparing their correlations to alternative measures of usability.

We propose using three qualitatively different measures to predict the user's task performance: past experience with similar interfaces, self-reported ratings of usability, and knowledge of how the interface works. Past experience with similar interfaces would indicate an ability to perform tasks more quickly and accurately on the new interface, because the user has already acquired relevant skills. Care must be taken in assessing experience to ensure that the questions are reliable and varied. Some studies have found that indices derived from questions about observable events tend to be more reliable than self-assessments of proficiency (i.e., questions of the form "In the past week, how many times have you used . . .?" instead of "Rate your proficiency with using . . ."; Kuniavsky, 2003; MacDorman, Vasudevan, & Ho, 2009).

Usability questionnaires typically ask people to rate how usable an interface is, for example, by asking them to state their level of agreement with various statements about a user interface (e.g., "I am satisfied with how easy it is to use this interface") and then accumulating the results into a usability index. These self-reported usability ratings provide an alternative measure of usability that is different from such performance measures as completion time and error rate but generally correlated with them nevertheless (Henderson, Podda, Smith, & Varela-Alvarez, 1995; Hornbæk & Law, 2007; Sauro & Lewis, 2009). There are a number of standardized usability questionnaires available (e.g., Bangor, Kortum,

& Miller, 2008; Brooke, 1996; Chin, Diehl, & Norman, 1988; Douglas, Kirkpatrick, & MacKenzie, 1999; Lewis, 1995, 2002).[1]

A test of knowledge about how the interface works is similar to a performance measure insofar as they both test knowledge about the interface. However, the former is a measure of declarative knowledge, whereas the latter reflects procedural knowledge (i.e., "knowing that" vs. "knowing how"; Anderson, 1976; Bargh & Chartrand, 1999; Dienes & Perner, 1999; Kirsh, 1991; Newell & Simon, 1972; Proctor & Van Zandt, 2008). Although user performance is often higher for procedural knowledge about interfaces than for declarative knowledge, both kinds of knowledge are correlated (Norman, 1988). Thus, users' degree of experience with similar interfaces, self-reported usability ratings of the interface, and declarative knowledge about how the interface works provide three different criteria for comparing measures by means of a correlational analysis. To automate study administration and minimize investigator influence in this study, questionnaires corresponding to these criteria are implemented at a Web site alongside an interface that records user performance data.

1.4. Research Questions

The time and the number of actions required to complete a step in a sequence of actions provide two objective sources of information with which to pinpoint usability problems in an interface. The purpose of this study is to determine which previously mentioned measure produces the highest construct validity in a usability measure derived from these two sources. In particular, the NEM is evaluated by comparing the construct validity of the ratio measure with that of alternative measures. Alternative measures proposed in the literature for group comparisons include the difference in group means, the difference in group medians, Cohen's d, and Hedges's $\hat{g}$. Any of these measures may be combined with logarithm or reciprocal transformations to normalize positively skewed data. For a large sample of participants interacting with a problematic interface, the method entails correlating each of these measures with the following usability correlates: experience with similar interfaces, subjective ratings of usability, and declarative knowledge about how the interface works:

RQ1: Which measure for comparing novice and expert group performance has the highest construct validity for measuring the usability of an interface based on completion time?

Completion time is only one measure of usability. Users are more likely to feel frustrated with an interface if its design causes them to make many errors. Completion times do not necessarily increase with error rate, especially when hastily performed actions can be quickly undone (e.g., using the backspace key when typing). Therefore, in addition to completion time, it is important to consider the number of actions taken to complete a step:

---

[1]Self-reported ratings also tend to be easier to interpret than usability measures based on user expectations (McGee, Rich, & Dumas, 2004).

RQ2: Which measure for comparing novice and expert group performance has the highest construct validity for measuring the usability of an interface based on the number of actions required to perform a task?

RQ3: Does a weighted combination of completion time and number of actions per step result in a single objective performance measure of usability that has higher construct validity than either completion time or number of actions per step considered separately?

2. METHODS

Care must be taken in developing an appropriate interface for this study. The features of the interface must vary sufficiently in their novelty and difficulty for the results to be generalizable, ensuring that the selected measure would work well for other interfaces. To obtain data with sufficient variance, the study should be performed on an interface that incorporates both familiar and counterintuitive features.

Few human–computer interfaces are as familiar as those of cellular (mobile) telephones. More than 4.1 billion cellular phone subscriptions are active worldwide (International Telecommunication Union, 2009). Mobile e-mail and other messaging services have become popular among businesspeople (Global System for Mobile Communications Association, 2006). Cellular phones are often used to send text messages using the simple messaging service. In December 2008, approximately 110 billion text messages were sent in the United States (International Association for the Wireless Telecommunications Industry, n.d.).

The test interface mimics iTap word prediction software for cellular phones but requires a two-key input method using a personal computer. Only about one third of the participants in this study had previously tried iTap; this level of inexperience is unsurprising, because most U.S. cellular phone handsets include T9 word prediction for text messaging instead of iTap. Participants with prior iTap experience should enjoy an advantage in using the interface owing to their existing mental models of how to use the commercial system (Norman, 1983). Hence, the interface was designed in such a way that most users could easily relate to what they were seeing on the screen, but the results of their actions would often defy their expectations. This was deemed important, because the interface must have usability problems to evaluate properly the construct validity of the measures in relation to variations in novice and expert group performance.

2.1. Participants

Data were collected and analyzed from 337 participants. Of those, 210 participants (62%) were female. The age group 18 to 22 years contained 228 participants (68%), and the maximum age was 57 years. The United States was the birthplace of 315 participants (94%). Participants were recruited by e-mail based on a random sample from a list of undergraduate students and recent graduates of eight campuses administered by a midwestern U.S. university.

2.2. Materials

After giving their informed consent, participants filled out a demographics questionnaire. The rest of the data collection instruments are listed next in the order in which they were administered:

- A questionnaire to assess automatic text completion experience, both in general and specifically using the iTap interface.
- Ten word-completion tasks to be performed using the iTap interface with a record of the completion time and number of clicks used.
- A questionnaire on interface usability adapted from the system usefulness factor of the IBM Computer System Usability Questionnaire (Lewis, 1995, 2002).
- A knowledge test of how the iTap interface works.

Participants were required to complete all data collection sections. The scores of each questionnaire and test were averaged separately to calculate the indices *completion experience, iTap experience, self-reported usability,* and *iTap knowledge.* The questions are listed in the appendix.

For the 10 word-completion tasks, participants were asked to spell a sequence of words: *phone, normal, comic, bench, pancake, enable, ring, tunnel, detach,* and *focus.* Words that vary the numerical patterns used in their spelling were selected to ensure that the time to complete the spelling would accurately reflect normal use of the interface and not be influenced by learning effects among words with similar spellings. A set of tasks that is more restricted than the tasks typically found in a usability study was chosen to reduce the variance of the performance data so that correlations between measures would reach statistical significance with fewer participants.

The completion time and number of clicks each participant used to spell a word correctly were recorded in the following manner: A click on any of the interface's buttons sent a page request to the computer server hosting the study Web site. The server recorded the times at which each page of the study was requested by participants and delivered by the server. Completion time for a given step was thus measured as the period between the server's delivery of the page and the participant's next click of a button. Clicks made outside the interface were ignored.

The interface follows a typical cellular phone layout (Figure 1). The box at the top displays the letters selected by the participant. The second, larger box displays the partial spelling of words. The rest of the interface is for input. As a number is selected on the keypad interface, the first letter associated with the number appears in the text box at the top. If this letter is the next letter in the word being typed (e.g., *a* for the *1-abc* key), the participant can simply select the next number. Otherwise, the participant must click on the appropriate partial spelling from a list. Once clicked, the selection will appear in the box at the top. This process continues until the word is spelled correctly. The participant then clicks submit to confirm that the spelling is correct and to continue to the next word. If the spelling is incorrect, the participant is prompted to fix it. A Delete button is provided for removing incorrect letters.

2.3. Procedures

All data collection instruments were administered at a Web site. Participants were first given a brief overview of the study and the technologies involved. After reading the study information sheet and giving their consent, participants provided their demographic data and filled out the automatic text completion experience and iTap experience questionnaires. Brief instructions and a sample interface were used to explain how to interact with the iTap interface. The participants were then presented with the first of the 10 words to spell (Figure 1). After successfully finishing the 10 word-completion tasks, they filled out the usability questionnaire and a test on knowledge about how the iTap interface works.

2.4. Statistical Analysis

Confirmatory factor analysis and Cronbach's alpha were used to evaluate the variability and internal reliability of the indices for automatic text completion experience, iTap experience, self-reported usability, and iTap knowledge. A correlation matrix was also calculated among these indices and completion time and number of clicks. Then the difference in group medians, the difference in the logarithm of group means, the ratio of group means, Cohen's d, and Hedges's $\hat{g}$ measures were appraised by correlating them with these four indices for task completion time and number of clicks taken.

Each of the 10 word-completion tasks produced two sets of novice–expert pairs. The following definitions eliminated variation in the expert group, which affords higher statistical significance than common alternatives (e.g., median splits, top
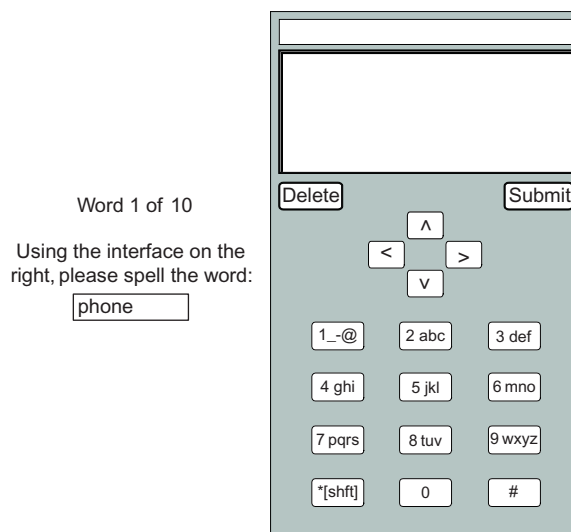

FIGURE 1   Word 1 as presented to the participant.

and bottom quartiles). For completion time, the expert group was defined as those participants who finished a particular word completion task in the shortest time achieved by any participant as measured in seconds, and the novice group was defined as the remaining participants. For number of clicks, the expert group was defined as those participants who used the minimum number of clicks for a particular word-completion task, and the novice group was defined as the remaining participants. Therefore, the expert group for completion time and the expert group for number of clicks varied from task to task in their size and composition and also differed from each other for any given task.

Table 1 lists the definition of each measure where t signifies task completion time, which is the time required to enter each word using the Web-based iTap interface. The same measures were used for the number of clicks taken to enter each word.

In this study, t_{expert} (or t_e) is defined to be the minimum task completion time in seconds achieved by any of the participants, and $clicks_{expert}$ is defined to be the minimum number of clicks required to complete the task. Also, t_{novice} is defined to be the average task completion time in seconds of novice participants, and $clicks_{novice}$ is defined to be the average number of clicks per task by the novice participants. The above variables represent averages for the 2nd through 10th word-completion task. The 1st task was excluded as a practice task. It should be noted that, in calculating Cohen's d and Hedges's $\hat{g}$, both SD (t_{expert}) and SD $(clicks_{expert})$ equal zero. The statistical significance used in this study is two-tailed with .05, .01, and .001 intervals.

The first task was excluded to reduce the effects of learning on the analysis of the performance data. It was anticipated that there would be a steep learning curve on the 1st word-completion task and then a leveling off on the 2nd through 10th word-completion tasks. (The results would later confirm that the learning effect was greater for the 1st task than for the 2nd through 10th tasks combined.) The interface knowledge test was placed after the 10 word-completion tasks, because this placement was expected to measure more accurately the user's knowledge during the nine scored tasks than a placement before the large learning effect of the 1st task.

Table 1:   Novice–Expert Performance Measure: Completion Time

Measure	*Equation*
Median	$\text{median}\,(t_{novice}) - \text{median}\,(t_{expert})$
Logarithm	$\text{mean}[\log(t_{novice})] - \text{mean}[\log\,(t_{expert})]$
Ratio	$\dfrac{[\text{mean}(t_{novice})}{\text{mean}(t_{expert})}$
Cohen's d	$\dfrac{\text{mean}(t_{novice}) - \text{mean}(t_{expert})}{\sqrt{\left\{[SD(t_{novice})]^2 + [SD(t_{expert})]^2\right\}/2}}$
Hedges's $\hat{g}$	$\dfrac{\text{mean}(t_n) - \text{mean}(t_e)}{\sqrt{\left\{(N_n-1)[SD(t_n)]^2 + (N_e-1)[SD(t_e)]^2\right\}/(N_n + N_e - 1)}} \times \left(\dfrac{3}{4(N_n+N_e)-9}\right)$

All tasks were simulated by KLM-GOMS in CogTool to confirm the validity of the t_{expert} and *clicks*$_{expert}$ estimates (Luo & John, 2005). Fortunately, no estimate exceeded the expert completion time predictions of CogTool using the minimum thinking time per step of 600 ms (Rogers & Monsell, 1995). The advantage of using actual participants instead of CogTool to estimate expert performance is that the novice–expert performance measures are more stable across the 10 word-completion tasks, because both novices and experts experience learning effects in sequence. If CogTool were used to simulate expert performance, novice performance would reflect learning effects but not the expert performance to which it was being compared.

3. RESULTS

This section uses confirmatory factor analysis to verify several distinct concepts that are correlated with novice–expert performance measures. To determine which measure has the highest construct validity, these factors are correlated with the median, logarithm, ratio, Cohen's d, and Hedges's $\hat{g}$ measures for average task completion time and average number of clicks. Finally, optimal weightings between task completion time and number of clicks are determined.

On average a word-completion task was finished in 20.28 s ($SD = 5.57$) and 11.11 clicks ($SD = 0.95$). On average each step required 1.82 s to complete ($SD = 0.47$). There were no significant gender or age differences in performance for completion time or number of clicks. The minimum time and number of clicks depended on the word completion task. The average minimum time for the nine scored tasks was 9.87 s, and the average minimum number of clicks was 9.78.

The participants who completed a task in the minimum time composed the expert group for completion time for that task. For the nine tasks, the expert group for completion time on average had 2 participants. The participants who used more time for a task composed the novice group for completion time for that task. For the nine tasks, the novice group for completion time on average had 335 participants ($M = 20.33$, $SD = 9.87$). The fastest novices were only 1 s slower than the experts. The participants who completed a task with the minimum number of clicks composed the expert group for number of clicks for that task. For the nine tasks, the expert group for number of clicks on average had 155 participants. The participants who made more clicks composed the novice group for number of clicks. For the nine tasks, the novice group for number of clicks on average had 182 participants ($M = 12.32$, $SD = 2.25$). The most effective novices required only one more click than the experts. The relative scarcity of experts, especially in terms of completion time, indicates that the new interface was not universally learnable (in contrast to a study involving only experts; e.g., Nielsen & Phillips, 2003). Applying logarithmic transformations to the two sets (time and clicks) did not improve their normality significantly.

For completion time, Cohen's d was 1.81 on average for Words 2 through 10 ($SD = .45$), and Hedges's $\hat{g}$ was 1.28 ($SD = 0.32$). For number of clicks, Cohen's d was 0.93 on average ($SD = 0.47$), and Hedges's $\hat{g}$ was .87 ($SD = 0.35$). In comparing the average estimates of Cohen's d and Hedges's $\hat{g}$, for completion time Cohen's

d overestimated the effect size by .53 s or 13%, $t(336) = 36.51$, $p = .000$, and for number of clicks Cohen's d overestimated the effect size by .06 or 41%, $t(336) = 9.690$, $p = .000$. These results show the improved accuracy of effect size estimates when applying Hedges's $\hat{g}$ to novice and expert groups.

3.1. Factor Analysis

Confirmatory factor analysis was used to test whether each of the experience, usability, and knowledge indices measures a single, unidimensional concept. If questions belonging to a single index fail to load on one factor, this indicates that the index may be measuring more than one concept. Factor analysis was applied to the combined questions of the three questionnaires: (a) the prestudy questionnaire of user experience with automatic text completion, (b) the poststudy usability questionnaire, and (c) the poststudy test of declarative knowledge about how the interface works (see the appendix). Table 2 shows the matrix used to transform the factor loadings in the component matrix. Table 3 shows that questions related to automatic text completion experience and iTap experience loaded on two different factors (Factor 2 and 4). However, the self-reported usability questions loaded on a single factor (Factor 1) as did the iTap knowledge questions (Factor 3). These results indicate that the indices for automatic text completion experience, iTap experience, self-reported usability, and knowledge about how the interface works all correspond to distinct and unidimensional concepts with respect to the interface in the study.

Next, factor analysis was applied to each questionnaire separately. Once again the experience-related questions loaded on two factors. The *completion experience* factor, which included both automatic text completion in general and specifically on a cellular phone, accounted for 57.18% of the variance ($s^2 = 3.43$). The *iTap experience* factor accounted for 23.50% of the variance ($s^2 = 1.41$). Together these factors accounted for 80.68% of the variance. Cronbach's alpha was .82 for completion experience, which shows high internal reliability.

Factor analysis was also applied to the poststudy usability questionnaire. All questions loaded on a single factor, which explained 62.13% of the variance ($s^2 = 4.97$). Cronbach's alpha was .91, which shows very high internal reliability for self-reported usability.

Table 2:   Component Transformation Matrix of Combined Questions From the Completion Experience and Usability Questionnaires and Test of Knowledge About How the Interface Works

	Factor			
Factor	*1*	*2*	*3*	*4*
1	.97	.12	−.19	.10
2	−.16	.88	−.03	.45
3	.18	.07	.98	−.01
4	−.03	−.46	.04	.89

Table 3:   Rotated Factor Matrix of Combined Questions From the Completion Experience and Usability Questionnaires and Test of Knowledge About How the Interface Works

| | | _Factor_ | | | |
		_1_	_2_	_3_	_4_
System Usability	I am able to enter a word quickly using this interface.	**.83**	−.02	−.09	−.03
System Usability	It was simple to use this interface.	**.83**	−.11	.00	.01
System Usability	I am satisfied with how easy it is to use this interface.	**.82**	−.05	−.09	−.04
System Usability	I can effectively enter a word using this interface.	**.82**	.04	−.05	.02
System Usability	I feel comfortable using this interface.	**.82**	.00	−.13	.07
System Usability	I believe I became productive quickly using this interface.	**.76**	.15	−.14	.03
System Usability	I am able to enter a word efficiently using this interface.	**.74**	.01	−.05	.03
System Usability	It was easy to learn to use this interface.	**.65**	.12	.05	.05
Completion Experience	Skill level of automatic text completion technology on a cellular phone	.03	**.89**	−.06	.14
Completion Experience	Skill level of automatic text completion technology on any kind of device	−.02	**.86**	−.01	.12
Completion Experience	Frequency using automatic text completion on any kind of device	.08	**.82**	.04	.06
Completion Experience	Frequency using automatic text completion on a cellular phone	.00	**.80**	−.02	.16
iTap Knowledge	If the next letter is _b_ and you press the button _2 abc,_ do you need to select _b_ before entering the next letter?	.00	−.05	**.73**	−.02
iTap Knowledge	How many mouse clicks are required to type _aaa_ and submit it?	−.13	.10	**.61**	−.16
iTap Knowledge	How many mouse clicks are required to type _cab_ and submit it?	−.16	−.02	**.60**	−.09
iTap Knowledge	If the next letter is _w_ and you press the button _9 wxyz,_ do you need to select _w_ before entering the next letter?	.01	−.13	**.51**	.22
iTap Knowledge	What requires more mouse clicks? Typing _add_ or _bad?_	−.12	.04	**.45**	.01
iTap Knowledge	If you are typing _phone_ and you have already typed _phm,_ only one choice appears: _pho,_ do you still need to select _pho_ before entering the next letter?	.07	.00	**.41**	.02
iTap Experience	Frequency of iTap on a cellular phone	.04	.24	−.01	**.94**
iTap Experience	Skill level using iTap on a cellular phone	.05	.26	−.03	**.92**

Note: Extraction method: Principal Component Analysis; Rotation method: Varimax with Kaiser Normalization. Rotation converged in five iterations.

Last, factor analysis was applied to the test of declarative knowledge about how the iTap interface works. The *iTap knowledge* factor accounted for 32.10% of the variance ($s^2 = 1.93$). Cronbach's alpha was .52 for iTap knowledge. An alpha value below .70 generally indicates subpar reliability.

3.2. Correlations Between Factors

Correlations were calculated between average completion time; average number of clicks; and the factors automatic text completion experience, iTap experience, self-reported usability, and iTap knowledge (Table 4). The results show that satisfaction with the usability of the interface was correlated with faster word entry ($r = -.18, p = .001$) and with fewer clicks ($r = -.24, p = .000$). Completion time was also correlated with number of clicks ($r = .23, p = .000$). These small-to-medium-sized correlations are typical of test-level correlations among effectiveness (e.g., errors), efficiency (e.g., time), and self-reported satisfaction measures according to a meta-analysis of the raw data from 73 usability studies (Hornbæk & Law, 2007).[2] In addition, participants who tested higher on declarative knowledge about how the iTap interface works completed words with fewer clicks ($r = -.27, p = .000$). Although participants with more automatic text completion experience were able to complete words faster, they made more errors. This increased error rate is likely to be caused by transfer effects in which acquired cognitive motor skills from other similar interfaces are applied to an interface with some procedural differences. (In other words, the participants tried to apply previously acquired procedural knowledge in using the new interface; however, the new interface violated familiar design conventions, causing the participants to make more errors.) The increased

Table 4: Correlation Matrix of Automatic Text Completion Experience, iTap Experience, Self-Reported Usability, iTap Knowledge, Clicks on Average, and Completion Time on Average

	Completion Experience	iTap Experience	Self-Reported Usability	iTap Knowledge	Average Clicks	Average Time
Completion experience	—					
iTap experience	.37***	—				
Self-reported usability	.05	.07	—			
iTap knowledge	−.05	.01	.01	—		
Average clicks	.11*	−.01	−.24***	−.27***	—	
Average time	−.11*	−.06	−.18**	−.10	.23***	—

**Correlation is significant at the .01 level, two-tailed.
***Correlation is significant at the .001 level, two-tailed.

---

[2]The test-level correlation between satisfaction and effectiveness or satisfaction and efficiency is typically low to medium when measured in the manner described in this study: only once for each participant after the completion of all performance testing. However, the task-level correlation between satisfaction and effectiveness or satisfaction and efficiency is much higher when measured after each task (Sauro & Lewis, 2009).

error rate indicates the interface may be counterintuitive, when viewed from the standpoint of participants' experience with existing interfaces.

3.3. Correlation of Novice-Expert Measures by Factor

The results indicate that the correlation between the four factors and average completion time and average number of clicks was similar for logarithm, ratio, Cohen's d, and Hedges's $\hat{g}$ (Table 5). For completion time, completion experience reached statistical significance for logarithm and Hedges's $\hat{g}$ measures only ($r = -.12, p = .024$). For number of clicks, self-reported usability had a slightly higher correlation with the ratio and Hedges's $\hat{g}$ measures ($r = -.29, p = .000$) than the logarithm measure ($r = -.27, p = .000$). The median measure's performance was poorer ($r = -.14, p = .011$). The results for Cohen's d were identical to Hedges's $\hat{g}$ with the exception that self-reported usability had a slightly smaller effect size ($r = -.27, p = .000$).

3.4. Correlation of Weighted NE Ratio Measure

The results of the NE ratio measure for average completion time and average number of clicks were converted to comparable units, namely, z scores, and a weighted average was calculated from these measures. The value of the weight was adjusted from 0 (100% completion time, 0% number of clicks) to 1 (0%, completion time, 100% number of clicks) at 10% increments.

Figure 2 shows the correlation between the weighted NE ratio measure and iTap experience, self-reported usability, and iTap knowledge. Cellular phone experience was excluded because of the sign change: People with more cellular phone experience actually clicked more, because they made more errors.

Table 5:   Correlation of Automatic Text Completion Experience, iTap Experience, Self-Reported Usability, and iTap Knowledge With the Median, Logarithm, Ratio, Cohen's *d*, and Hedges's *ĝ* Measures by Completion Time and Number of Clicks

	Median	Log	Ratio	Cohen's d	Hedges's ĝ
			Completion Time		
Completion experience	−.11*	−.12*	−.10	−.12*	−.12*
iTap experience	−.09	−.07	−.06	−.06	−.06
Self-reported usability	−.16*	−.17**	−.20**	−.20**	−.20**
iTap knowledge	−.05	−.08	−.09	−.10	−.10
			No. of Clicks		
Completion experience	.05	.09	.10	.08	.08
iTap experience	−.07	−.02	−.01	−.02	−.02
Self-reported usability	−.14*	−.27**	−.29**	−.27**	−.29**
iTap knowledge	−.26**	−.27**	−.27**	−.26**	−.26**

*Correlation is significant at the .05 level, two-tailed.
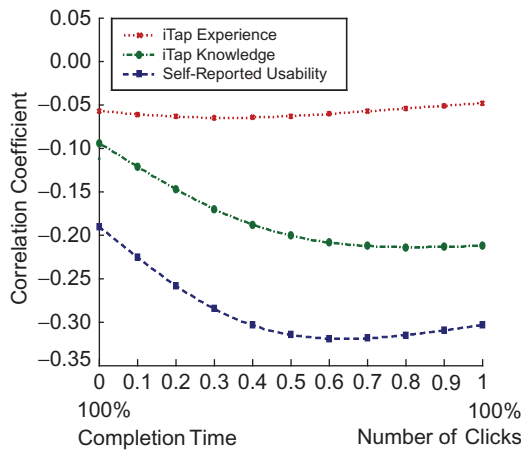**Correlation is significant at the .01 level, two-tailed.

FIGURE 2 A weighted average of the novice–expert ratio measure for average completion time in *z* scores and average number of clicks in *z* scores is correlated with iTap experience, self-reported usability, and iTap knowledge.

Table 6:   Optimal Correlation and Weight of Self-Reported Usability, iTap Knowledge, and iTap Experience with Hedges's *ĝ*, Cohen's *d*, Logarithm, Ratio, and Median Measures

	Self-Reported Usability		*iTap Knowledge*		*iTap Experience*	
Measure Rank	*Correlation*	*Weight*	*Correlation*	*Weight*	*Correlation*	*Weight*
Hedges's *ĝ*	−.32***	0.65	−.26***	0.86	−.06	0.20
Cohen's *d*	−.31***	0.69	−.27***	0.89	−.06	0.24
Logarithm	−.30***	0.85	−.28***	0.95	−.06	0.39
Ratio	−.32***	0.64	−.21***	0.81	−.06	0.30
Median	−.26***	0.96	−.21***	0.80	−.05	0.77

***Correlation is significant at the .001 level, two-tailed.

Table 6 lists the correlations and optimal weights for the Hedges's *ĝ*, Cohen's *d*, logarithm, ratio, and median measures. For the Hedges's *ĝ* measure, self-reported usability reaches the highest inverse correlation ($r = −.32$, $p = .000$) at an optimal weight of 0.65, iTap knowledge reaches the second highest inverse correlation ($r = −.26$, $p = .000$) at an optimal weight of 0.86, and iTap experience reaches the lowest inverse correlation ($r = −.06$, $p = .257$) among the three at an optimal weight of 0.20. At the optimal weights, the construct validity for Hedges's *ĝ* was marginally higher than for Cohen's *d* and logarithm. They were followed by ratio and median measures. It is worth noting that although the correlation between the weighted Hedges's *ĝ* and self-reported usability was only medium sized ($r = −.32$, $p = .000$), the correlation was much higher than correlations typically found in the literature between satisfaction measures like self-reported usability and error rate ($r = −.196$, 95% confidence interval $= ±.184$) or satisfaction and completion time ($r = −.145$, 95% confidence interval $= ±.129$; see Table 5 of Hornbæk & Law, 2007).

4. DISCUSSION

For measuring the usability of an interface based on completion time (RQ1) and based on the number of actions required to perform a task (RQ2), the logarithm, ratio, Cohen's d, and Hedges's $\hat{g}$ measures had somewhat similar construct validity. In addition, all four measures had higher construct validity than the median measure for both completion time and number of actions. However, for this data set, Hedges's $\hat{g}$ provided a significantly more accurate measure of effect size than the other measures, and it provided a more stable measure of subsamples.

Thus, in answer to the first two research questions, although there are no strong empirical grounds for selecting a measure based solely on construct validity, there are empirical grounds for selecting Hedges's $\hat{g}$ based on effect size accuracy. In addition, Cohen's d and Hedges's $\hat{g}$ have the benefit of being standard statistical measures of effect size that can be applied to other types of data. This can make it easier to compare data on completion time and number of actions taken with other types of data. Hedges's $\hat{g}$ has the additional advantage of providing an accurate estimate even when the sample sizes and standard deviations of the novice and expert groups differ. Nevertheless, the ratio measure has the heuristic benefit of being easy to explain (e.g., "The novices took on average x times longer than the experts").

The NEM only uses step completion time to identify usability problems. It ignores the number of actions taken to complete a step. However, the results show that self-reported usability was more strongly correlated with number of actions taken ($r = -.27$ to $-.29$, $p = .000$) than completion time ($r = -.17$ to $-.20$, $p = .002$ to .000) for the logarithm, ratio, Cohen's d, and Hedges's $\hat{g}$ measures (Table 5). The results also show that iTap knowledge was more strongly correlated with number of actions taken ($r = -.26$ to $-.27$, $p = .000$) than completion time ($r = -.08$ to $-.10$, $p = .139$ to .057). The significance was strong for number of actions, but .05-level significance was not reached for completion time. For completion time, only the experience-related factors were more strongly correlated with the four top performing measures (i.e., Hedges's $\hat{g}$, Cohen's d, logarithm, and ratio); only automatic text completion experience reached significance for any of these measures, and it failed to reach significance for the ratio measure. In sum, these results indicate that the four best measures, including the NE ratio measure, have higher construct validity for number of actions taken than for completion time. This indicates that methods that compare novice and expert performance to assess usability should include measures derived from both kinds of information.

Figure 2 shows that, for the NE ratio measure, a weighted average of completion time and number of actions taken has higher construct validity than either NE ratio by itself (R3). This trend held for the logarithm, Cohen's d, and Hedges's $\hat{g}$ measures. Hedges's $\hat{g}$ outperformed the ratio measure for iTap knowledge and tied it for self-reported usability and iTap experience. For the ratio measure, the optimal weight was 0.30 for iTap experience, 0.64 for self-reported usability, and 0.81 for iTap knowledge. For Hedges's $\hat{g}$, the optimal weight was 0.20 for iTap experience, 0.65 for self-reported usability, and 0.86 for iTap knowledge. Given the strength and significance of the inverse correlation for self-reported usability,

its corresponding weight seems like a reasonable value to use for the interface in this study. Thus, in answer to the third research question, it seems advisable to combine step completion time and number of actions per step into a single objective performance measure of usability.

5. CONCLUSION

The main contribution of this study is to devise and validate a usability measure based on novice and expert performance

- that is a much better predictor of self-reported usability than previously published methods, and
- whose accuracy is not biased by variations in the size and performance of the novice and expert groups.

The −.32 correlation between the self-reported usability index and the weighted Hedges's $\hat{g}$, which combines data for both step completion time and number of actions per step, is much larger in magnitude than correlations between satisfaction and completion time ($r = -.145$) or satisfaction and number of actions taken ($r = -.196$) in other studies that also employed only a single posttest usability questionnaire (Hornbæk & Law, 2007; Sauro & Lewis, 2009). The −.32 correlation is also nearly double the −.18 correlation obtained in the study by using the original NEM (Urokohara et al., 2000).

In the current study, the ratio measure was compared with some common alternative measures, such as Cohen's d and Hedges's $\hat{g}$. Except for the median, which fared poorly, the candidate measures had somewhat similar construct validity. Overall, Hedges's $\hat{g}$ provided only marginally higher construct validity than Cohen's d. The main justification for using Hedges's $\hat{g}$ is analytical. It is the only measure that provides comparable effect size estimates across studies that include expert and novice groups that differ in size and in the standard deviation of their performance data as was the case in this study. In addition, Hedges's $\hat{g}$ allows for the accurate comparison of different kinds of data.

For all measures tested, the results show that the highest incremental construct validity was obtained by calculating a weighted average of the NE value for step completion time and for number of actions taken. This summary measure is more strongly correlated with self-reported usability than either NE value individually. The current study's results show that self-reported usability was more strongly inversely correlated with number of actions taken than with step completion time. The magnitude of this correlation only increased when applying the NE ratio measure to number of actions taken. This supports the observation made in the introduction that a user may feel unhappy with making errors even when the errors can be quickly corrected.

This study has practical implications for the applied human–computer interaction (HCI) practitioner. The study has devised and validated a new usability measure with the following benefits:

- The measure is objective and performance based, avoiding the apparent subjectivity of cognitive walk-throughs, think-aloud protocols, and heuristic evaluation and their low interevaluator reliability (Hertzum & Jacobsen, 2003). Heuristic evaluation, for example, can lead different HCI practitioners to arrive at different results for the same interface, because many principles must be weighted in making an analysis (Faulkner & Wick, 2005).
- Nevertheless, the measure is correlated with self-reported usability with very high statistical significance and a medium effect size.
- The measurement of user performance can easily be embedded in a Web-based interface to facilitate rapid and convenient recruitment, data collection, and analysis.

These benefits indicate that the new measure could be a useful tool for usability engineers to diagnose aspects of interfaces that make them difficult for novices to use.

One limitation of the new measure and the other evaluated measures is that they indicate only at what step in a task a usability problem appears. This identifies *what* the usability problem is and *where* it occurred, but it does not indicate *why* the user is having the problem and *how* to fix it. However, once a problem has been identified, HCI practitioners can fall back on traditional methods, such as heuristic evaluation, user interviews, and focus groups, to answer the *why* and *how* questions.

The authors expect the weighted novice-expert Hedges's $\hat{g}$ method to generalize to other interfaces that meet the following criteria:

- The interface belongs to an application for which users have well-defined goals that they are trying to achieve efficiently and effectively.
- The design intention is to create an interface that is intuitive for novices to use, and not necessarily one that is optimized for expert performance.

An important limitation to note is that the best interface for the job is not necessarily the interface that is easiest for novices to use. For example, a telemarketer working on commission may prefer an interface with a steep learning curve, if tasks can be performed very fast after sufficient practice (Nielsen, 1994). Powerful applications like Maya 3D animation software target professionals who demand speed and functionality, which necessarily makes the applications challenging for novices. Thus, the measures evaluated in this study are useful when the user's first impression of the interface is of primary concern.

REFERENCES

Anderson, J. R. (1976). *Language, memory and thought*. Hillsdale, NJ: Erlbaum.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, *24*, 574–594.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, *54*, 462–479.

Bessiere, K., Newhagen, J. E., Robinson, J. P., & Shneiderman, B. (2006). A model for computer frustration: The role of instrumental and dispositional factors on incident, session, and post-session frustration and mood. *Computers in Human Behavior*, *22*, 941–961.

Brinkman, W.-P., Haakma, R., & Bouwhuis, D. G. (2007). Towards an empirical method of efficiency testing of system parts: A methodological study. *Interacting with Computers*, *19*, 342–356.

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189–194). London: Taylor and Francis.

Carroll, J. M., & Rosson, M. B. (1987). Paradox of the active user. In J. M. Carroll (Ed.), *Interfacing thought: Cognitive aspects of human-computer interaction* (pp. 80–111). Cambridge, MA: MIT Press.

Chin, J. P., Diehl, V. A., & Norman, K. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the Sixth ACM/SIGCHI Conference on Human Factors in Computing Systems* (pp. 213–218). New York: ACM Press.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.

Desurvire, H., Lawrence, D., & Atwood, M. (1991). Empiricism versus judgement: Comparing user interface evaluation methods on a new telephone-based interface. *SIGCHI Bulletin*, *23*(4), 58–59.

Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences*, *22*, 735–808.

Dillon, A., & Song, M. (1997). An empirical comparison of the usability for novice and expert searchers of a textual and a graphical interface to an art-resource database. *Journal of Digital Information*, *1*(1), Article No. 2.

Douglas, S. A., Kirkpatrick, A. E., & MacKenzie, I. S. (1999). Testing pointing device performance and user assessment with the ISO 9241, Part 9 Standard. In *Proceedings of the 17th ACM/SIGCHI Conference on Human Factors in Computing Systems* (pp. 215–222). New York: ACM Press.

Faulkner, L., & Wick, D. (2005). Cross-user analysis: Benefits of skill level comparison in usability testing. *Interacting with Computers*, *17*, 773–786.

Global System for Mobile Communications Association. (2006, November 28). *Strong global demand for MMS and mobile email*. Retrieved June 29, 2009, from http://www.gsmworld.com/newsroom/press-releases/2118.htm

Goonetilleke, R. S., Shih, H. M., On, H. K., & Fritsch, J. (2001). Effects of training and representational characteristics in icon design. *International Journal of Human-Computer Studies*, *55*, 741–760.

Gould, J., & Lewis, C. (1985). Designing for usability: Key principles and what designers think. *Communications of the ACM*, *28*, 300–311.

Gray, W. D., & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human–Computer Interaction*, *13*, 203–261.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*, 197–216.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Henderson, R., Podda, J., Smith, M., & Varela-Alvarez, H. (1995). An examination of four user-based software evaluation methods. *Interacting with Computers*, *7*, 412–432.

Hertzum, M. & Jacobsen, N. E. (2003). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, *15*(1), 183–204.

Hornbæk, K., & Law, E. L.-C. (2007). Meta-analysis of correlations among usability measures. *Proceedings of the ACM/SIGCHI Conference on Human Factors in Computing Systems* (pp. 617–626). New York: ACM Press.

International Association for the Wireless Telecommunications Industry. (n.d.). *Wireless quick facts*. Washington, DC: Author. Retrieved on June 29, 2009, from http://www.ctia.org/media/industry_info/index.cfm/AID/10323

ISO 9241-11. (1997). *Ergonomic requirements for office work with visual display terminals (VDTs), Part 11: Guidance on usability specification and measures*. Geneva, Switzerland: International Organization for Standards.

International Telecommunication Union. (2009). *Measuring the information society: The ICT development index*. Geneva, Switzerland: Author.

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, 320–351.

Kang, N. E., & Yoon, W. C. (2008). Age- and experience-related user behavior differences in the use of complicated electronic devices. *International Journal of Human-Computer Studies*, 66, 425–437.

Kirsh, D. (1991). When is information explicitly represented? In P. Hanson (Ed.), *Information, thought, and content* (pp. 340–365). Vancouver, Canada: UBC Press.

Kjeldskov, J., Skov, M. B., & Stage, J. (2005). Does time heal? A longitudinal study of usability. In S. Balbo & T. Bentley (Eds.), *Proceedings of the Australian Computer-Human Interaction Conference*. Narrabundah: CHISIG of Australia.

Kuniavsky, M. (2003). *Observing the user experience: A practitioner's guide to user research*. San Francisco: Morgan-Kaufmann.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human–Computer Interaction*, 7(1), 57–78.

Lewis, J. R. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human–Computer Interaction*, 14, 463–488.

Lindgaard, G., & Dudek, C. (2003). What is this evasive beast we call user satisfaction? *Interacting with Computers*, 15, 429–452.

Luo, L., & John, B. (2005). Predicting task execution time on handheld devices using the keystroke-level model. In *Extended Abstracts of the 23rd ACM/SIGCHI Conference on Human Factors in Computing Systems* (pp. 1605–1608). New York: ACM Press.

MacDorman, K. F., Vasudevan, S. K., & Ho, C.-C. (2009). Does Japan really have robot mania? Comparing attitudes by implicit and explicit measures. *AI & Society*, 23, 485–510.

MacKenzie, I. S. (1992). Fitts' Law as a research and design tool in human-computer interaction. *Human–Computer Interaction*, 7(1), 91–139.

McGee, M., Rich, A., & Dumas, J. (2004). Understanding the usability construct: User-perceived usability. In *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting* (Vol. 48, pp. 907–911). Santa Monica, CA: HFES.

Molich, R., & Nielsen, J. (1990). Improving a human–computer dialogue. *Communications of the ACM*, 33, 338–348.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

Nielsen, J. (1992). Finding usability problems through heuristic evaluation. *Proceedings of the 10th ACM/SIGCHI Conference on Human Factors in Computing Systems* (pp. 373–380). New York: ACM Press.

Nielsen, J. (1994). *Usability engineering*. San Francisco: Morgan Kaufmann.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the Eighth ACM/SIGCHI Conference on Human Factors in Computing Systems* (pp. 249–256). New York: ACM Press.

Nielsen, J., & Phillips, V. L. (1993). Estimating the relative usability of two interfaces: Heuristic, formal, and empirical methods compared. In *Proceedings of the INTERACT'93 and CHI'93: The 11th ACM/SIGCHI Conference on Human Factors in Computing Systems* (pp. 214–221). New York: ACM Press.

Norman, D. A. (1983). Some observations on mental models. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 7–14). Hillsdale, NJ: Erlbaum.

Norman, D. A. (1988). *The psychology of everyday things*. New York: Basic Books.

Proctor, R. W., & Van Zandt, T. (2008). *Human factors in simple and complex systems* (2nd ed.). Boca Raton, FL: CRC Press.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*, 510–532.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*, 207–231.

Sauro, J., & Lewis, J. R. (2009). Correlations among prototypical usability metrics: Evidence for the construct of usability. In *Proceedings of the 27th International ACM/SIGCHI Conference on Human Factors in Computing Systems*. New York: ACM Press.

Silfverberg, M., MacKenzie, I. S., & Korhonen, P. (2000). Predicting text entry speed on mobile phones. In *Proceedings of the ACM/SIGCHI Conference on Human Factors in Computing Systems* (pp. 9–16). New York: ACM Press.

Uehling, D. L., & Wolf, K. (1995). User action graphing effort (UsAGE). In *Conference Companion on Human Factors in Computing Systems* (pp. 290–291). New York: ACM Press.

Urokohara, H., Tanaka, K., Furuta, K., Honda, M., & Kurosu, M. (2000). NEM: "Novice expert ratio method" A usability evaluation method to generate a new performance measure. In *Proceedings of the ACM/SIGCHI Conference on Human Factors in Computing Systems* (pp. 185–186). New York: ACM Press.

APPENDIX

Prestudy Completion Experience Questionnaire

1. How would you rate your expertise using automatic text completion technology on any kind of device?
   No Experience, Beginner, Competent, Proficient, Expert
2. In the past week how many times have you used automatic text completion on any kind of device (e.g., computer, including web address and search term completion, cell phone, PDA)?
   0, 1, 2, 3, 4, 5+
3. How would you rate your expertise using automatic text completion technology on a cellular phone?
   No Experience, Beginner, Competent, Proficient, Expert
4. In the past year how many times have you used automatic text completion on a cellular phone?
   0, 1, 2, 3, 4, 5+
5. How would you rate your expertise using the iTap interface on a cellular phone?
   No Experience, Beginner, Competent, Proficient, Expert

6. How many times have you ever used the iTap interface on a cellular phone?
   0, 1, 2, 3, 4, 5+

Poststudy Usability Questionnaire (adapted from Lewis, 1995, 2002)

1. Overall, I am satisfied with how easy it is to use this interface.
   Strongly Disagree, Moderately Disagree, Slightly Disagree, Neutral, Slightly Agree, Moderately Agree, Strongly Agree
2. It was simple to use this interface.
3. I can effectively enter a word using this interface.
4. I am able to enter a word quickly using this interface.
5. I am able to enter a word efficiently using this interface.
6. I feel comfortable using this interface.
7. It was easy to learn to use this interface.
8. I believe I became productive quickly using this interface.

Poststudy Test of Knowledge about How the Interface Works

1. If the next letter is *w* and you press the button *9 wxyz,* do you need to select *w* before entering the next letter?
   Yes, No, Don't know
2. If the next letter is *b* and you press the button *2 abc,* do you need to select *b* before entering the next letter?
   Yes, No, Don't know
3. If you are typing *phone* and you have already typed *phm,* only one choice appears: *pho;* do you still need to select *pho* before entering the next letter?
   Yes, No, Don't know
4. What requires more mouse clicks? Typing *add* or *bad?*
   add, bad, They both take the same number of steps, Don't know
5. How many mouse clicks are required to type *aaa* and submit it?
   3, 4, 5, 6, 7, Don't know
6. How many mouse clicks are required to type *cab* and submit it?
   3, 4, 5, 6, 7, Don't know