Meta-Analysis of Correlations Among Usability Measures

Kasper Hornbæk
Department of Computer Science
University of Copenhagen
Universitetsparken 1, 2100 Copenhagen, Denmark
kash@diku.dk

Effie Lai-Chong Law
Computer Engineering and Networks Laboratory
Eidgenössische Technische Hochschule Zürich
Gloriastrasse 35, CH-8902, Zürich, Switzerland
law@tik.ee.ethz.ch

ABSTRACT

Understanding the relation between usability measures seems crucial to deepen our conception of usability and to select the right measures for usability studies. We present a meta-analysis of correlations among usability measures calculated from the raw data of 73 studies. Correlations are generally low: effectiveness measures (e.g., errors) and efficiency measures (e.g., time) have a correlation of .247 ± .059 (Pearson's product-moment correlation with 95% confidence interval), efficiency and satisfaction (e.g., preference) one of .196 ± .064, and effectiveness and satisfaction one of .164 ± .062. Changes in task complexity do not influence these correlations, but use of more complex measures attenuates them. Standard questionnaires for measuring satisfaction appear more reliable than homegrown ones. Measures of users' perceptions of phenomena are generally not correlated with objective measures of the phenomena. Implications for how to measure usability are drawn and common models of usability are criticized.

Author Keywords

Usability evaluation, usability measures, ISO 9241, subjective satisfaction, meta analysis

ACM Classification Keywords

H.5.2 [Information Interfaces and Presentation (e.g., HCI)]: User Interfaces—Evaluation/Methodology; D.2.2 [Software Engineering]: Design Tools and Techniques—User Interfaces

INTRODUCTION

Usability is commonly understood as a broad notion indicating the quality-in-use of interactive systems [e.g., 3,15]. Following this understanding, measures of usability are plentiful and diverse, and include task completion time, error rates, subjective satisfaction, perceived workload, assessments of a work product's quality, feelings of enjoyment, questionnaires on ease-of-use, and so forth.

Such measures of usability play important roles in several areas of human-computer interaction: usability engineering has as its goal to use usability measures to improve computer systems; research comparing the relative merits of two interfaces often uses measures such as task completion times and errors; and practical summative testing of an application against a competitor's product also typically relies on usability measures. While the quantitative measures of usability we discuss in this paper are not the only way to capture the usability of an interface, they are widely used and indispensable to many researchers and practitioners.

The literature on HCI, however, offers surprisingly little help in how to measure usability, in particular how to select measures of usability. The papers investigating this issue have mostly looked at correlations between usability measures, but show mixed results [9,16,20,23]. Nielsen and Levy [23], for example, found that performance and preference were correlated in 75% of a selection of 57 studies, meaning that users in general preferred the application with which they performed best. In contrast, Frøkjær et al. [9] argued that the usability aspects of effectiveness, efficiency, and satisfaction should be measured independently and not in general be expected to correlate. In addition to addressing these differences in results, it has been suggested that analysis of correlations among usability measures would help understand better how usability can be measured [13].

We present a meta-analysis of usability measures by investigating how they correlate in 73 studies. The aim of the analysis is to provide information about how measures relate, which will help (1) understand better what usability is and how to develop models of it, and (2) select measures for usability studies. In contrast to earlier meta-analyses of usability we use the raw data of the studies, allowing calculations to be thorough and uniform across studies; we base our results on a comprehensive sample of journals and conferences, forming the largest sample we know of to be meta-analyzed with respect to usability; we investigate the role of moderator variables such as task complexity; and we present implications for both usability research and practical usability studies.

CHI 2007, April 28–May 3, 2007, San Jose, California, USA.
Copyright 2007 ACM 978-1-59593-593-9/07/0004...$5.00.

RELATED WORK

The literature on usability contains numerous definitions of usability and models of the dimensions or components of usability [e.g., 15,26,27,28]. Shneiderman and Plaisant [28], for example, identified five usability measures: time to learn, speed of performance, rate of errors by users, retention over time, and subjective satisfaction. The ISO 9241-11 standard [15] identified three aspects of usability: effectiveness, efficiency and satisfaction. Seffah et al. [26] developed a synthesizing model of usability based on existing work. Their QUIM model incorporates more than 127 specific measures in 10 factors, including – in addition to the ISO aspects – factors such as safety, trustfulness and accessibility. The main contribution from this line of work appears to be its fleshing out the meaning of the usability construct and its implications for how to measure usability.

In practice, however, choosing among measures appears difficult. A recent review of usability measures used in HCI research listed more than 54 kinds of measure [13]. This diversity – and the desire to develop empirically based models of usability – has spurred studies of the extent to which usability measures are related. Typically this is done by studying the correlations between usability measures [e.g., 2,9,17,23]. Below we briefly review these studies.

Nielsen and Levy [23] performed a meta-analysis of 57 papers, investigating the relation between objective and subjective measures of performance. They used published papers as their source of data, from which information about usability measures was extracted. This information was analyzed so as to uncover whether objective performance measures and subjective preference measures showed similar results, that is, favored the same interface. Nielsen and Levy found that in approximately three-quarter of the cases, performance predicted preference.

In contrast, Bailey [2] presented an early argument for separating measures of preference and performance. Similarly, Kissel [17] found that subjective and objective measures of usability were only weakly related, but that this relation was affected by users' experience with computers. A study by Frøkjær et al. [9] also challenged Nielsen and Levy's conclusions. In an analysis of data from a single experiment, and from a selection of papers from the ACM's CHI conference, they found no correlation between the three aspects of usability identified by the ISO 9241-11 standard. They recommended measuring all the three aspects - effectiveness, efficiency, and satisfaction.

However, recently papers have appeared that try to combine usability measures, in part on the assumption that they to some degree contribute the same information. McGee [20] and Sauro and Kindlund [16] shared the goal of developing a single, standardized usability score. McGee derived his usability score from participants' subjective assessments of tasks (i.e., usability magnitude estimation), whereas Sauro and Kindlund computed theirs by summating the values of four objective and subjective usability measures (i.e., completion, time, error, and satisfaction). These single scores have the apparent advantage of brevity. However, the validity of McGee's master usability scale, with sole reliance on user perception, is constrained by how users interpret the definition of usability. For instance, they may selectively and inconsistently focus on certain aspects when assessing the usability of an object. Similarly, the validity of Sauro and Kindlund's summated usability score is limited by which usability metrics that are included in or excluded from their summation procedure. For their approach to work, we still need to establish which aspects or metrics of usability that are valid.

A relatively recent and separate approach to understanding how usability measures are related departs from users' perception of product qualities and their relation. McGee et al. [21], for example, conducted a study of how users weighted 64 potential usability characteristics. They used multivariate analysis to arrive at five groups of usability aspects, including core usability (e.g., clear and easy-to-learn), secondary usability (e.g., helpful and accessible), and satisfaction qualities (e.g., attractive and interesting). While studies such as that of McGee et al. certainly enrich our understanding of usability, it is not a priori obvious that users' understanding of usability is the whole or even a main component of usability.

In summary, correlation studies appear one of the most prominent sources for understanding usability and how it may be measured. Yet, the findings of correlation studies are in contradiction and limited on at least four counts. First, existing studies (with the exception of the study by Nielsen and Levy [23]) base their conclusions on a limited number of data sets. Second, papers that study correlations often do not have access to raw data, only summary statistics. Already Nielsen and Levy [23] described how they could not use common techniques for meta-analysis because "the original papers did not report sufficient statistical detail about their results" (p. 69). Thus, relations between usability aspects can only be simplistically coded. Third, studies of correlations rarely account for the variety of ways that, for instance, satisfaction may be measured and what this means for relations between usability measures. Fourth, studies of correlations do not try to account for contextual factors (e.g., task complexity) that may impact the relation between usability aspects. The aim of the meta-analysis presented next is to address these limitations.

METHOD OF META-ANALYSIS

The goal of the meta-analysis is to study the relation among usability measures, using the raw data from a selection of published studies. In particular we aim to investigate (a) the relation between usability aspects such as effectiveness, efficiency, and satisfaction; (b) the relation between specific measures, for example, task completion time and NASA's Task Load Index [11]; (c) the variables that may moderate these relationships, such as task complexity and the use of particular types of measures; and (d) the implications of the relations identified for usability research and practical usability evaluation.

The overall phases of the meta-analysis are to select studies for inclusion, to attempt obtaining raw data for these studies, to code the studies, and to analyze the coded studies. Below we go through these phases. First, however, we outline the basic procedures and goals of meta-analysis.

Procedures and goals of meta-analysis

In general, meta-analysis is an organized way to summarize, integrate and interpret selected sets of empirical studies. Through systematic procedures of coding, recording and computing, effects and relationships on which a set of studies converge (or diverge) can be identified. The most important concept supporting these activities is that of effect size, a quantification of the magnitude of a difference between conditions or of a relation among variables. Effect size is related to the significance of a statistical test so that *significance = number of subjects x effect size*. This relation makes it possible to combine effect sizes that may not be significant in individual studies to form a general and possibly significant picture of some phenomena of interest. We base our work on the existing literature on meta-analysis, specifically the procedures of Glass et al. [10], Rosenthal [25], Lipsey and Wilson [19], and Hunter and Schmidt [14].

In this study meta-analysis is a matter of aggregating the correlations of usability measures across studies, because correlations are one way of expressing an effect size (in fact the preferred one of for example [25]). As will be discussed below, the main difference between typical meta analyses, including meta-analyses in HCI [e.g., 30], and our study, is that we have the raw data of studies available.

Selection of candidate studies

As candidates for inclusion in the meta-analysis we consider studies from eight HCI journals and conferences, see Table 1. We chose these sources because they represent a broad spectrum of work in HCI. We looked at studies from the years 2003 through 2005. This range was chosen because it yielded a substantial number of full papers for consideration (2090) and because we expected it to become increasingly difficult to get in contact with authors of the papers if we chose a longer span of time.

As candidates for our analysis we focus on *original research papers reporting usability measures concerning human interaction with user interfaces*. Let us expand on this focus. First, we only looked at full-length papers reporting original research; we assumed that short papers, poster summaries, and session overviews would not contain the kind of detail needed to perform the meta-analysis.

Second, since the meta-analysis concerns correlations between measures, a candidate study had to report at least two measures. We disregarded studies with no information on usability measures (e.g., because the studies were formative or preliminary) and studies reporting only qualitative data. The latter choice aimed to restrict the focus of the meta-analysis; it does not imply that we find studies of a qualitative nature (e.g., reports of usability problems) of lesser utility in HCI.

Third, because our focus is on human performance we excluded studies that did not have this as their primary focus. Thus, studies of cognitive models were excluded, as were papers concerned with testing data collection methods or with exploring specific sociological or psychological research questions.

Fourth, because the focus is on interaction we excluded studies with no two-way exchange of information between the user and the computer; some studies of non-interactive reading to compare the legibility of different font sizes, for example, were excluded on this account.

Fifth and finally, we interpreted user interfaces somewhat narrowly in that we disregarded interfaces in support of driving and flying/aviation. This was done to ensure a relatively homogeneous sample, to use somewhat similar domains, and to limit the number of papers from the HFES

Source	*Full papers*	*Candidates*	*Included with raw data*
ACM Conference on Human Factors in Computing Systems	261	94	29
ACM Transactions on Computer-Human Interaction	49	16	4
Annual Meeting of the Human Factors and Ergonomics Society	1165	99	14
Behaviour & Information Technology	99	30	6
Human-Computer Interaction	36	5	1
IFIP TC13 International Conference on Human-Computer Interaction	152	49	10
International Journal on Human-Computer Studies	201	64	7
Interacting with Computers	127	29	2
Total	2090	386	73

Table 1. Studies included in the meta-analysis distributed over sources (years 2003-2005).

conference (which, as Table 1 suggests, is disproportionally large compared to the other sources of studies).

Overall, we considered as candidate studies 386 (18%) of the full papers published in the sources.

Obtaining raw data from candidate studies

Existing meta-analyses of usability measures have not had access to the raw data of studies; this is how meta-analyses are typically carried out, see [10,14,19,25]. When sufficient information is contained in the papers being analyzed, this is fine. However, correlations between usability measures are typically *not* reported in the HCI literature. Thus, only coarse coding of the dependent variables is possible. This was, for example, what Nielsen and Levy [23] did, by noting whether performance and preference data suggested the same direction of differences between conditions. However, we wanted to have the raw data of the studies available so that we could calculate correlations ourselves and, more importantly, so that we could quantify the effect size of relations among usability aspects.

To obtain the raw data, we contacted authors of the 386 candidate studies to inquire if they would share with us the original data of their studies (or a random subset of the data if they were more comfortable with that). The first author was contacted by e-mail; if we received no reply we followed up with an e-mail to the author that we perceived to be the senior researcher; in some cases we also mailed a letter to the address mentioned in the paper.

This procedure yielded responses from 184 authors; an overall response rate of 48%. Among the authors who responded, 133 (35%) agreed to share their data; 92 (24%) of these authors actually sent in their datasets. Fifty-one of the authors (13%) declined our requests for data sharing for various reasons: being prohibited from sharing because of institutional review boards or ethics guidelines, data loss, no time to retrieve data, data currently under use, or no access to data.

Some authors sent us more than one data set, for example when a paper reported several experiments. In those cases, we randomly chose only one of the data sets, as inclusion of both studies could bias our sample.

Nineteen of the 92 sent-in datasets were discarded because of incompleteness, inappropriate data format or unclear research methodologies. Consequently, we processed and analyzed 73 sets of raw data, that is, 19% of the total number of candidate studies.

Coding studies and dependent variables

Studies were coded in part on their methodology (e.g., number of participants, study design, duration of tasks) and in part on their so-called substantive dimensions (e.g., the domain). Two codes require explanation. Task complexity was coded on a three-item rating scale: low, medium, and high; we used Rasmussen's [24] work to inform this scale, so that low complexity tasks were skill based (e.g., clicking on objects), medium complexity tasks were rule based (e.g., navigating an information space), and high complexity tasks were knowledge based (e.g., drafting privacy policies). Domain was coded using the leaf-levels of the ACM Computing Classification System (http://www.acm.org/class/1998/).

After extracting from individual studies all their dependent variables, each variable was classified according to ISO 9241-11 standard – the tripartition of usability into effectiveness, efficiency, and satisfaction [15]. Then, each variable was further classified using a taxonomy from a recent study of usability measures [13]. This taxonomy distinguishes 54 kinds of usability measures. For example, asking study-participants to rank interfaces in terms of preference would first be categorized as regarding the ISO category of satisfaction, and next be classified as rank preferred interface, in the preference category of [13].

In several candidate studies, the same specific usability aspect is measured in several ways. For example, ease of use may be measured by a series of questions on a post-task questionnaire. Following [13], we code measures that are described as capturing the same construct as just one measure (e.g., trust or feelings of frustration). Standardized questionnaires [e.g., 4,11] are considered one measure, independently of the actual number or phrasing of questions. Note that in contrast to [13], we coded TLX as a standardized questionnaire.

To ensure that studies were reliably coded, both authors went over the coding of every study and resolved any differences in opinion by discussion and by consulting the paper and the raw data set.

Method of meta-analysis

After coding studies, we analyzed the raw data from individual studies and then aggregated across the studies. The meta-analysis is barebones, not using most of the many corrections and models available [14]; we hope that this makes the paper more generally accessible.

Analysis of raw data

For each raw data file we calculated correlations between the usability measures. Note that correlations may be calculated at different levels, for example, at the task level for individual subjects, at the level of averages for a particular subject's measures for a specific condition, or at the level of averages for a particular interface. We calculate correlations at the least-aggregated level possible. When computing the correlation between two measures such as time and accuracy, the correlation will be based on the time and accuracy data for each subject's solution to each task. Some measures are not available at the task level, but only per interface (a frequent example is subjective satisfaction measures). Correlations will then be calculated per subject per interface, that is, on each subject's average task completion time and satisfaction score for an interface. In all cases, we carefully checked the sign of the correlation.

For checking whether conditions in the studies impacted the correlations, we used the residual correlations from multivariate analysis of variance.

As suggested by Glass et al. [10, p. 148] all correlations were transformed into Pearson's product-moment correlation coefficient (r). Procedures for doing so are readily available in the meta-analysis literature [14]. In particular, error and preference are often dichotomous, meaning relations between them and other variables will be point-biserial correlations, which need transformation. The product-moment correlations will be used as our effect size measure, with the usual interpretation that r^2 signifies the variance explained (or how well one variable predicts another) and that an $r \approx .5$ is a large effect, $r \approx .3$ is a medium effect, and $r \approx .1$ is a small effect [5].

Aggregation across studies

When aggregating effect sizes across studies, we first transform effect sizes to standard values using Fisher's r to z conversion [25]. The z-transformed score has a standard error of *1/sqrt (n-3)*, where n is number of participants in the study. The inverse of this error can be used as a weight for each individual z-transformed score, so that studies with smaller standard errors are given more emphasis. In practice, this amounts to multiplying the effect size with *n-3*. After this weighting, studies can be aggregated by averaging their z-transformed scores: Rosenthal [25] suggests this as a conservative procedure. Finally, z-transformed scores can be translated back to r values, which is what we report throughout the paper.

RESULTS

We first give an overview of the characteristics of the studies in our sample and how they measure usability. Next, we discuss the correlations among measures. Note that throughout the paper, effectiveness, efficiency, and satisfaction will be reported with the meaning that higher values are better: high accuracy and low error rates are thus both indicative of high effectiveness.

Descriptive data on the selection of studies

Altogether 73 datasets were analyzed. The domain of the studies was categorized according to the ACM classification system (see Table 2).

The average number of tasks performed by each participant per study was 104 (*SD* = 251, ranging from 1 to 1440). Correspondingly, the mean task duration was 226 seconds (*SD* = 437, ranging from 1 to 1860 seconds). On average, a study lasted 0.56 hours (*SD* = 1.31, ranging from about 2 minutes to 10 hours). Ten of the 73 studies did not present data on task duration. In some studies a task was a simple move-and-click with a mouse, in others a single task could consist of many complex sub-steps or last for days. In terms of task complexity [24], 34 studies were low complexity (perceptual motor, e.g., click-and-pick), 25 were medium complexity (rule-based cognitive, e.g. search), and 14 were high complexity (problem-solving, e.g., route planning).

Domain	*Frequency*
Input devices and strategies	16
Information presentation and navigation	14
Information search and retrieval	10
Interaction device, style and technique	8
Graphical user interface	7
Visualization	4
Virtual environment	4
Audio-based interaction	2
Database management	2
Distributed collaborative computing	2
Evaluation/Methodology	2
Others (programming, trust)	2

Table 2. Distribution of domains.

The average number of participants involved per study was 32 (*SD* = 29, ranging from 6 to 181). In 37 studies the participants were experienced with respect to the tasks required to be performed, and in 23 studies the participants were novice. Two studies employed both types of subject, whereas 11 studies did not give any data in this regard. The studies' research designs were directly related to the number of participants recruited: 44 studies employed within-subject repeated design (mean N = 19); 23 studies between-subject (mean N = 51); two studies mixed design (mean N = 25), and in four studies the design was unknown.

Measures used

The measures taken in the 73 studies were categorized according to ISO 9241-11 and to the taxonomy developed by Hornbæk [13]. Note that a usability aspect, say efficiency, can be gauged by different measure *types* (e.g., time) and their subsuming measure *tokens* (e.g., task completion time or time to event) [13]. A measure token (say error rate) can be gauged differently depending on the specific tasks performed in a study. Measure tokens of satisfaction, in particular, are difficult to classify because they are collected using a variety of questionnaires, scales and levels of granularity, which seem only bounded by the imagination of their authors.

Counted at the level of measure tokens, the overall average was 4.07 measures per study (*SD* = 1.85, range 2 to 9). Table 3 shows the corresponding values for the three usability aspects.

As shown in Figure 1, 36 out of the 73 studies (49%) had measures of all the three usability aspects; 30 studies (42%) had measures of the combination of effectiveness-efficiency, effectiveness-satisfaction or efficiency-satisfaction; seven of the studies (9%) collected measures of only one usability aspect. In some studies, the same

Measure	*M*	*SD*	*Range*
Effectiveness	1.18	0.86	0 – 4
Efficiency	1.33	0.90	0 – 4
Satisfaction	1.53	1.53	0 – 7

Table 3: Mean number of tokens for the usability aspects.

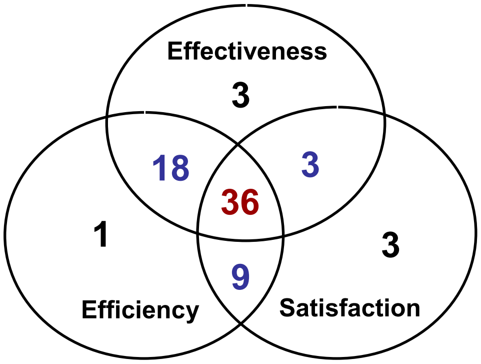

Figure 1. Venn diagram illustrating the number of studies measuring effectiveness, efficiency, and/or satisfaction.

measure token was collected several times. For instance, in a study where the task was multidirectional point-and-click, two types of error (number of overshoots and number of counterproductive movements) were registered; in this case, we counted the measure token – error rate – only once.

As shown in Figure 2, the distribution of the nine types of effectiveness can be represented by an exponential curve with the peak token (i.e., error rate, 35 instances, 47% of the studies reviewed included this measure) being followed by a series of less frequent tokens (e.g., spatial accuracy). The same distribution can be seen for the measure tokens of efficiency (Figure 3), where the peak token is task completion time (76% of the studies reviewed), followed by several specific tokens such as deviation from optimal path and percentage of preferred walking speed (under “Others”). These findings indicate that some convergence concerning selection of usability measures exists.

Studies contain a variety of satisfaction measures. Twenty-five measure tokens were identified when enumerating them at the finest (or third) level of the taxonomy from [13]. When they were grouped into the coarsest (or first) level, there were six groups (see Table 4).

Surprisingly, only 12 out of 106 instances of satisfaction measures (11%) employed standard questionnaires (QUIS [4]; ASQ, CSUQ [18]; NASA TLX [11]; Douglas et al.’s questionnaire [8]). The most popular measure type is “satisfaction with the interface”, which can be further broken into two measure tokens): ease-of-use (24 measures) and context-dependent-questions (10 measures). In contrast, the measure type “specific attitudes towards the interface” is more diverse, including annoyance, confidence, control, discomfort, frustration, fun, learnability, liking, and want-to-use-again. Further, the measure type “others” include tokens that are emerging (e.g., trust and beauty), vaguely defined (e.g., responsiveness) or encompassing (e.g., some sub-attributes of quality-in-use described in ISO 9126).

Satisfaction	*Frequency*
Satisfaction with the interface	34
Specific attitudes towards the interface	22
Users' attitudes and perceptions	16
Preference	13
Standard questionnaire	12
Others	9

Table 4. Distribution of satisfaction measures.

Correlation between effectiveness and efficiency

Figure 4 shows the average correlations between effectiveness and efficiency. Across the 54 studies that include measures of both these aspects we find a correlation of .247 (confidence interval, CI95%, ± .059). This suggests that more efficient performance, such as faster task completion, is associated with more effective performance, such as fewer errors. According to Cohen, this is a small to medium effect. In practical terms, 87% of the studies have a positive correlation between effectiveness and efficiency. Among the studies with the highest correlations is a study of four navigation interfaces for a hand-held mobile device (*r* = .79 between getting lost and time to complete the navigation); a study of authoring of privacy rules shows a negative correlation between errors and time usage, *r* = -.23. Though the average correlation is significantly above zero (as indicated by the confidence interval), it appears that there is quite a bit of variation in the data.

To uncover the sources of this variation, we may look at just the prototypical measures of effectiveness and efficiency, that is, at task completion rates and task completion time. Table 5 summarizes the average

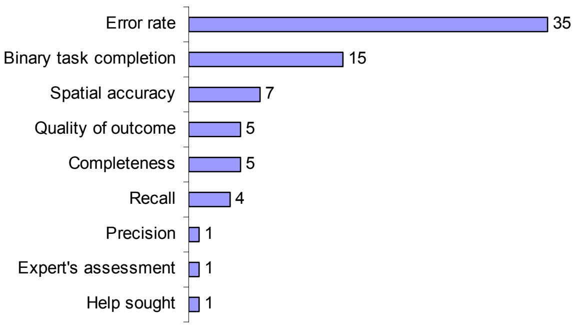

Figure 2. Distribution of effectiveness tokens.

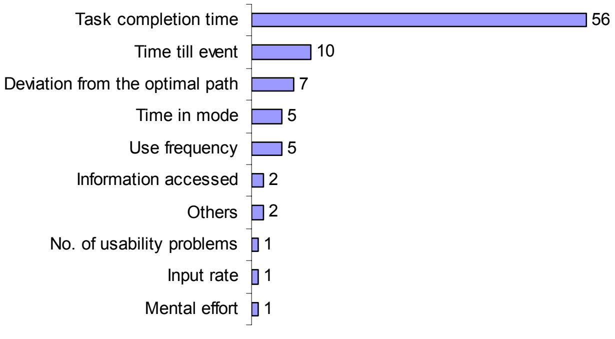

Figure 3. Distribution of efficiency tokens.

correlations for those measures, and suggests that the average correlation between those measures range from .145 to .316. Thus, the general result of a correlation between effectiveness and efficiency appears not to be due to our inclusion of a broad range of usability measures. Rather, the prototypical correlation is higher between time and error than the one presented above.

Task complexity does not seem to affect the relation between efficiency and effectiveness. Figure 4 illustrates the relationship between variables across the task complexity categories. It could be hypothesized, as done by for example [9], that more complex tasks would not show strong correlations between usability aspects. For none of the three combinations of ISO aspects of usability is task complexity significant (Hedge's test for homogeneity of variances, *Qs* = 0.58, 0.79, 1.89, all *ps* > .25). Thus, task complexity does not seem to attenuate or otherwise affect the relation between usability aspects.

Simple explanations related to task complexity seem difficult because the precise way effectiveness and efficiency are measured impacts their relationship. Take as one example the difference in what errors are taken to mean. Two interpretations may be found in the data: *errors-along-the-way* and *task-completion-errors*. Errors-along-the-way are mistaken actions on the way to task completion: trying a wrong navigation path or miss clicks before hitting an object; task-completion-errors are errors in a task's outcome (e.g., introducing new bugs in a code editing task, poor grades for essays written with computer support). Errors-along-the-way has an average correlation to efficiency of .441 ± .125 (14 studies); task-completion-errors an average correlation to efficiency of .155 ± .08 (23 studies). One reason behind this difference between correlations is that in many of the studies, especially those using low complexity tasks, making errors along the way negatively affected task time. This is the case, for example, in studies of input devices and interaction techniques that allowed only a correct selection of an object to end a task: missing the object would of course make the task last longer.

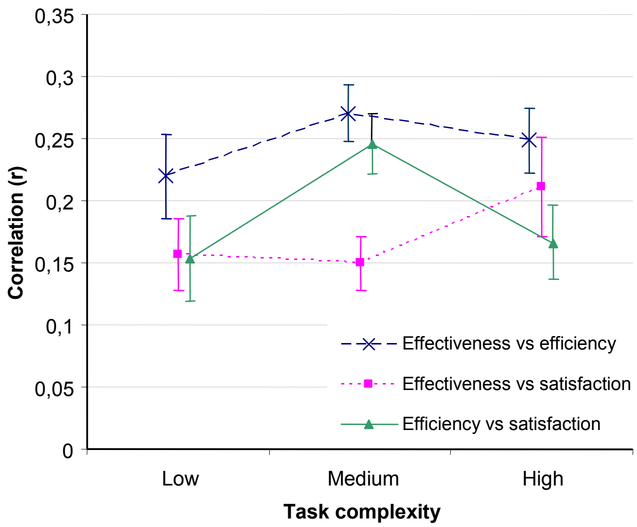

Figure 4. Correlations between measures of effectiveness, efficiency, and satisfaction. Effect sizes are Pearson's product-moment correlations; error bars give the standard error of the mean. Assessment of task complexity is based on [24].

	Errors	*Time*	*Satisfaction*
Errors			
Time	.316 (±.070)		
Satisfaction	.196 (±.184)	.145 (±.129)	
Preference	.243 (±.158)	.309 (±.146)	.245 (±.281)

Table 5: Correlations between prototypical measures of usability. The 95% confidence intervals are given in parentheses. Cells contain from 7 to 35 studies, except the preference-satisfaction correlation based on just two studies

In addition to the above explanation, it seems that some of the more complex measures of effectiveness are not correlated to efficiency measures. As suggested by Figure 2, most studies measure error rates and binary task completion; fewer assess quality (five studies) or use expert assessments (one study). None of these six studies show a significant correlation between efficiency and effectiveness. Rather, the average correlation between efficiency and complex satisfaction measures is negative (r = -.039). While task complexity does not in itself change relations between usability measures (the above studies are but one of high task complexity), complexity of measures does.

Correlation between effectiveness and satisfaction

Figure 4 also shows the average correlation between measures of effectiveness and satisfaction. Across the 39 studies that include measures of both these aspects we find a correlation of .164 ± .062. According to Cohen, this is a small effect. It is also the lowest of the three comparisons between ISO aspects. Yet, 86% of the studies show a positive correlation between effectiveness and satisfaction.

The simplest example of these relations occurs between on the one hand task completion rates and satisfaction questionnaires and preference indications on the other hand. Table 5 suggests correlations of .243 between error and preference and of .196 between errors and satisfaction questions. Since preference is typically measured dichotomously, we may illustrate the difference concretely by the observation that for studies in Table 5, error rates are about 18% for the non-preferred interfaces (or interfaces) and 13% for the preferred ones.

Six studies measure both task effectiveness (say accuracy) and participants' own assessment of their effectiveness. This happens, for instance, when post-task questionnaires are used to ask participants about their confidence in task solutions (e.g., "were your answers to tasks: very good –

very poor", "how do you perceive the accuracy of the technique just used: poor – very good"). The correlation between effectiveness (typically error rates or binary task completion) and participants' assessment of their task solutions is on average $r = .22$ (ranging from -.23 to .80), not significantly different from a correlation of zero. This observed inconsistency between objective and subjective measures could be due to cognitive and social bias, for example, the role of prior experience [29] and social desirability effect (e.g., [1]). This remains to be explored.

Correlation between efficiency and satisfaction

Figure 4 also shows the relation between efficiency and satisfaction. Across the 45 studies that report one or more measure of these aspects, the average correlation is .196 ± .064; a small to medium effect. However, it appears relatively uniform across studies as 81% of them show positive correlations. Again we can illustrate this correlation by appealing to the relation between prototypical measures, in this case between task completion times and preference/satisfaction questionnaires (see Table 5). For these studies, a preferred interface is about 20% faster than a non-preferred one.

As suggested earlier, a number of studies measure both efficiency and participants' experience of the interaction. This is done by questionnaire items like "rate your satisfaction with task completion time" and "how quickly did the system let you finish your tasks". Interestingly, we find a correlation between such questions and objective task completion times that are indistinguishable from zero (on average $r = .30$ across five studies). Again, correlations vary a lot, with two studies showing negative relations between time and subjective measures of the interaction.

Measures of satisfaction

A large portion of the studies uses several ways of measuring satisfaction, for example, by asking questions concerning specific satisfaction, using a standardized questionnaire, and asking for the interface that users preferred. This opens the possibility of studying the relation among measures of satisfaction which we do next.

One issue that the raw data of the studies allow us to investigate is the reliability of satisfaction measures. For all questionnaires where we had available the full questionnaire data, we calculated Cronbach's α, a widely used measure of the reliability of questionnaires [6]. Table 6 gives these values for standard questionnaires and for homegrown ones, that is, questionnaires created ad hoc with more than one question and that purport to measure some aspect of satisfaction. The table suggests that homegrown questionnaires have lower reliability and greater variation in reliability: six such questionnaires fail to reach the commonly accepted minimum reliability of .70. Possibly some of the questionnaires attempt to measure several distinct constructs, but we suspect the drop in reliability may be caused by poor questionnaire design.

Questionnaire	*N*	*Cronbach's α*	
		Mean	Range
Standard questionnaires (e.g., TLX, QUIS, CSUQ)	16	.814	.73 - .95
Homegrown	20	.736	.21 - .92

Table 6: Reliability of satisfaction questionnaires as indicated by Cronbach's α [6]. Homegrown refers to questionnaires that authors themselves developed to capture ease-of-use.

Three studies in our sample measure satisfaction both at the level of an individual task and at an aggregated level, typically once for each interface. For instance, in one study the simple three-question ASQ [18] was administered right after each task to capture users' instant reactions and the long 19-question CSUQ was administered after all the tasks had been performed to capture users' overall perception of the system. Reassuringly, the correlations between individual and task level satisfaction measures are medium to large, with *r*s ranging from .38 to .70.

Finally, 10 studies measure both preference and some other aspect of satisfaction. It appears relevant to look at how well satisfaction questionnaires filled out during a study predict the preferences expressed by participants. Again the correlation is reassuringly large, with a mean *r* of .49.

DISCUSSION

We have characterized how usability is measured across a selection of 73 studies. Our study has shown an overall small to medium correlation between usability aspects; typical measures of usability are related with a Pearson correlation coefficient ranging from .164 to .247. Factors involved in shaping these correlations are the use of complex usability measures, of prototypical or standardized measures, and of measures based on participants' perceptions. Task complexity does not seem to influence the relation. We find quite similar correlations across studies with clear differences in domains, interface types, procedures, and experimental conditions, suggesting that these differences matter less for relations among usability measures than commonly assumed.

Interpretations of the results

On the one hand our data may be interpreted as showing only a low correlation between usability aspects (what could be called the half empty interpretation). Though the effect sizes are generally low to medium, they might be considered of little practical importance given the variation in the data. This interpretation follows the literature that suggests weak to no correlation among usability measures [2,9,17]. Indeed the results that more complex measures of effectiveness and efficiency are not correlated, and that task-completion-errors attenuate correlations, suggest that in more complex study setups, correlations drop.

Under this interpretation, our analyses indicate that attempts to reduce usability to one measure [e.g., 16,20] are bound to lose important information, because there is no strong correlation among usability aspects. Apart from masking potentially interesting details, the use of a single usability score cannot support formative evaluation well. Further, data redundancy is not necessarily undesirable: even when usability measures are highly correlated they convey information in different ways. In system redesign, for example, developers may be more convinced or motivated to improve the system when usability measures converge.

The correlations found in our paper are lower than those presented by Sauro and Kindlund [16]. They found correlations between time and error of r = .5, and between satisfaction and task completion of about .5. Possible reasons for their overestimation relative to our data include simple measures used consistently across studies (rather than the realistic variety of measures we have studied), a much smaller data set (they had 129 participants, our data sets cover more than 2000), a specific kind of task (as opposed to the variety we studied), and the use of per-task satisfaction measures (as opposed to administering questionnaires only once or twice per participant as is the general case in our dataset). With correlations half the size of those in Sauro and Kindlund's study, we think the argument behind the one-measure usability score is seriously weakened.

Our other interpretation (the half-full interpretation) is reflecting a surprise to see such a general correlation across studies. In about 80-90% of the studies, variables in the main categories of the ISO classification are positively correlated. The straightforward idea by Frøkjær et al. [9] that as task complexity increases the relation between usability aspects decreases does not seem to hold. Rather, an important factor in attenuating correlations is more complex measures of in particular effectiveness: these include quality of tasks, task-completion-errors, and so on.

The complexity of usability evaluation tasks are somewhat tied to the domain. Our finding that task complexity does not affect the relationships among usability measures seems to imply that such measures are applicable across a wide spectrum of domains. Further, the "ceiling" or "floor" effect [20] engendered by the particularities of tasks that are non-canonical, but tailor-made by usability specialists for evaluating a particular system, does therefore not appear threatening to the utility of task-based performance metrics.

Implications for usability research

Our study suggests several conceptual problems in current models of usability. First, our distinction between errors-along-the-way and task-completion-errors indicates a particular relation between these measures and other usability measures. We do not see this distinction in common models of usability (e.g., ISO 9241), nor do recommendations on selecting measures for usability tests make this distinction (or describe its implications). Further, error identification is not a clear-cut process in certain situations as evaluators may diverge on what constitutes an error [16], especially when the cause of the error is considered (e.g., a slip or a mistake). Analogously, we find inconsistencies in classifying workload. It is measured similarly to satisfaction measures and correlate strongly with such measures, but some authors [e.g., 13] consider it an efficiency measure

Second, we find a difference between users' experience of interaction/outcomes and objective measures. For studies that collect both measures of the same phenomenon, we find negligible correlations. While some models accept fundamental differences between subjective and objective measures [13,31], others do not [23]. Leveraging these differences for novel measures would be interesting. Czerwinski et al. [7] suggested the relation between perceived time assessment and actual time passed as a novel usability measure. This idea might be extended to measures other than time.

Third, the variation in usability measures suggests the malleability and extensibility of the notion of usability. Among others, the user experience movement [12] has argued to broaden the notion of usability, rather than narrowing it. We find mixed results related to this issue, because some aspects of users' experience seem orthogonal to performance measures and some shows substantial correlations. Further work may investigate if correlational studies could help describe the relation between user experience indicators and traditional usability measures.

Implications for studies of usability

We suggest that usability studies describe correlations among the usability measures collected. This would help interpret and compare outcomes of usability evaluations. We also recommend that standard questionnaires be used when possible, given their higher reliability, and that the more complex effectiveness measures be used when feasible (as they are more likely to give information that cannot be obtained by measures in the other categories).

Meta-analytic caveats and open questions

Our method raises a couple of concerns. First, we have extensively relied upon the ISO classification of usability. As mentioned above we do not find it entirely satisfactory, but it has helped manage the complexity of our data set. Second, in a few of the studies in our sample there are differences between the correlations found when aggregating data at the task level and at the user interface level; we currently cannot offer any good explanation for these differences. Third, our meta-analysis has been barebones. We consider using path modeling to investigate the relation among usability aspects promising, and would like to further analyze the role of continuous moderator variables, such as duration of use. Fourth, it is not our wish to scorn arguments that usability is to a large extent shaped by context [e.g., by 22]. There are a large number of variables whose significance we have not investigated. This leads to perhaps the most pressing open research question, namely the lack of useful predictive theories about the relations between usability aspects.

CONCLUSION

Studies of correlations among usability aspects appear a useful way of enriching our understanding of usability. However, existing studies are in disagreement with each other and often calculate correlations from a limited collection of data. We have investigated correlations in the raw data of 73 usability studies and find medium to low correlation among usability measures. In addition a number of specific factors that affect correlations have been identified (e.g., complex measures, subjective versus objective measures). Our results suggest that some models of usability are problematic and that theory to speculate about the relation between measures is lacking.

ACKNOWLEDGEMENTS

We are unusually grateful to the many authors that provided data and positive responses to our many requests.

REFERENCES

1. Arnold, H. J. & Feldman, D. C. Social Desirability Response Bias in Self-Report Choice Situations, *Academy of Management Journal*, *24* (1981), 377-385.
2. Bailey, R. Performance vs. Preference, *Proc. HFES 1993*, 282-286.
3. Bevan, N. Measuring Usability As Quality of Use, *Software Quality Journal*, *4* (1995), 115-150.
4. Chin, J. P., Diehl, V. A., & Norman, K. L. Development of an Instrument for Measuring User Satisfaction of the Human-Computer Interface, *Proc. CHI '88*, 213-218.
5. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences,* Lawrence Erlbaum Associates, 1969.
6. Cronbach, L. J. Coefficient Alpha and the Internal Structure of Tests, *Psychometrika*, *16* (1951), 297-333.
7. Czerwinski, M., Horvitz, E., & Cutrell, E. Subjective Duration Assessment: An Implicit Probe for Software Usability? *Proc. IHM-HCI 2001*, (2001), 167-170.
8. Douglas, S. A., Kirkpatrick, A. E., & MacKenzie, I. S. Testing Pointing Device Performance and User Assessment with the ISO 9241, Part 9 Standard, *Proc. CHI'99*, 215-222.
9. Frøkjær, E., Hertzum, M., & Hornbæk, K. Measuring Usability: Are Effectiveness, Efficiency, and Satisfaction Really Correlated? *CHI2000*, 345-352.
10. Glass, G., McGaw, B., & Lee Smith, M. *Meta-Analysis in Social Research*, Sage Publications, 1981.
11. Hart S. & Staveland, L., Development of NASA-TLX, in Hancock & Meshkati *Human Mental Workload,* Elsevier, 1988, 139-183.
12. Hassenzahl, M. & Tractinsky, N. User Experience – a Research Agenda, *Behaviour & Information Technology*, *25*, 2 (2006), 91-99.
13. Hornbæk, K. Current Practice in Measuring Usability: Challenges to Usability Studies and Research, *International Journal of Human-Computer Studies*, *64*, 2 (2006), 79-102.
14. Hunter, J. & Schmidt, F. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, Sage Publications, Thousand Oaks, CA, 2004.
15. ISO Ergonomic Requirements for Office Work With Visual Display Terminals (VDTs)-Part 11: Guidance on Usability, 1998.
16. Kindlund, E. & Sauro, J. A Method to Standardize Usability Metrics into a Single Score, *Proc. CHI 2005*, 401-409.
17. Kissel, G. The Effect of Computer Experience on Subject and Objective Software Usability Measures, *Proc. CHI'95*, (1995), 284-285.
18. Lewis, J. IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use, *International Journal of Human-Computer Interaction*, *7*, 1 (1995), 57-78.
19. Lipsey, M. & Wilson, D. *Practical Meta-Analysis*, Sage Publications, Thousand Oaks, CA, 2000.
20. McGee, M. Master Usability Scaling: Magnitude Estimation and Master Scaling Applied to Usability Measurement, *Proc. CHI 2004*, 335-342.
21. McGee, M., Rich, A., & Dumas, J. Understanding the Usability Construct: User-Perceived Usability, *Proc. HFES'2004*, 907-911.
22. Newman, W. & Taylor, A. Towards a Methodology Employing Critical Parameters to Deliver Performance Improvements in Interactive Systems, *Proc. INTERACT'99*, 605-612.
23. Nielsen, J. & Levy, J. Measuring Usability: Preference Vs. Performance, *Comm. ACM*, *37*, 4 (1994), 66-75.
24. Rasmussen, J. Skills, Rules, and Knowledge: Signals, Signs, and Symbols, and Other Distinctions in Human Performance Models, *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-13*, 3 (1983), 257-266.
25. Rosenthal, R. *Meta-Analytic Procedures for Social Research*, Sage Publications, Newbury Park, CA, 1991.
26. Seffah, A., Donyaee, M., Kline, R., & Padda, H. Usability Measurement and Metrics: A Consolidated Model, *Software Quality Journal*, *14* (2006), 159-178.
27. Shackel B., Usability - Context, Framework, Definition, Design and Evaluation, in *Human Factors for Informatics Usability,* 1991, 21-38.
28. Shneiderman, B. & Plaisant, C. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, Addison Wesley, Boston, MA, 2005.
29. Taylor, S. & Todd, P. Assessing IT Usage: the Role of Prior Experience, *MIS Quarterly*, *Dec.* (1995), 561-570.
30. Weisband, S. & Kiesler, S. Self Disclosure on Computer Forms: Meta-Analysis and Implications, *Proc. CHI 1996*, 3-10.
31. Yeh, Y.-Y. & Wickens, C. D. Dissociation of Performance and Subjective Measures of Workload, *Human Factors*, *30*, 1 (1988), 111-120.