Available online at www.sciencedirect.com
SCIENCE DIRECT®
Int. J. Human-Computer Studies 64 (2006) 79–102
International Journal of Human-Computer Studies
www.elsevier.com/locate/ijhcs

Current practice in measuring usability: Challenges to usability studies and research

Kasper Hornbæk*

Department of Computer Science, University of Copenhagen, Universitetsparken 1, DK-2100 Copenhagen, Denmark

Received 23 February 2004; received in revised form 27 May 2005; accepted 14 June 2005
Available online 15 August 2005
Communicated by M. Atwood

Abstract

How to measure usability is an important question in HCI research and user interface evaluation. We review current practice in measuring usability by categorizing and discussing usability measures from 180 studies published in core HCI journals and proceedings. The discussion distinguish several problems with the measures, including whether they actually measure usability, if they cover usability broadly, how they are reasoned about, and if they meet recommendations on how to measure usability. In many studies, the choice of and reasoning about usability measures fall short of a valid and reliable account of usability as quality-in-use of the user interface being studied. Based on the review, we discuss challenges for studies of usability and for research into how to measure usability. The challenges are to distinguish and empirically compare subjective and objective measures of usability; to focus on developing and employing measures of learning and retention; to study long-term use and usability; to extend measures of satisfaction beyond post-use questionnaires; to validate and standardize the host of subjective satisfaction questionnaires used; to study correlations between usability measures as a means for validation; and to use both micro and macro tasks and corresponding measures of usability. In conclusion, we argue that increased attention to the problems identified and challenges discussed may strengthen studies of usability and usability research.

Keywords: Usability; User-centered design; Usability Engineering

1. Introduction

The purpose of this paper is to review current practice in how usability is measured, and to analyse problems with the measures of usability employed. On that basis, we discuss challenges to conducting usability studies and to research into how to measure usability.

Usability is a core term in human–computer interaction (HCI). Among the efforts to explain what the term means, usability has been called "the capability to be used by humans easily and effectively" (Shackel, 1991, p. 24); "quality in use" (Bevan, 1995); and "the effectiveness, efficiency, and satisfaction with which specified users can achieve goals in particular environments" (ISO, 1998, p. 2). Most explanations of what usability means agree that it is context dependent (Newman and Taylor, 1999) and shaped by the interaction between tools, problems and people (Naur, 1965). A key research question in HCI is how to work with and improve the usability of interactive systems. Research addressing this question has led to guidelines for improving the usability of systems (Smith and Mosier, 1986), methods for predicting usability problems (Molich and Nielsen, 1990; Wharton et al., 1994), techniques to test the usability of systems (Lewis, 1982), and discussions on how to measure usability (Nielsen and Levy, 1994; ISO, 1998; Frøkjær et al., 2000). This paper takes as its point of departure the latter question.

Our focus on how to measure usability has three motivations. First, what we mean by the term usability is to a large extent determined by how we measure it. Explanations in textbooks on HCI of what usability means

*Tel.: +45 35321425; fax: +45 35321401.
E-mail address: kash@diku.dk.

1071-5819/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijhcs.2005.06.002

are often made in terms of aspects of the use of a computer system that can be measured (Nielsen, 1993, p. 26; Dix et al., 1993, pp. 131–145). Thus, measures of usability serve to make the general and somewhat vague term usability concrete and manageable.

Second, usability cannot be directly measured. Through operationalization of the usability construct, we find aspects of usability that can be measured. The choice of such measures not only fleshes out what usability means, it also raises the question if that which is measured is a valid indicator of usability. The difficulty of finding valid measures is well described in the literature on the validity of psychological constructs (Cook and Campbell, 1979; American Psychological Association, 1985). Attention to usability measures may thus uncover validity problems in how usability is operationalized and reasoned about.

Third, many approaches to user-centered design depend critically on measures of the quality of interactive systems, for example by benchmarking against usability measured for previous versions or competitors' systems (Gould and Lewis, 1985; Whiteside et al., 1988). Similarly, the goal of usability engineering (Nielsen, 1993) of working quantitatively with usability also rests on measures of usability. The question of which measures of usability to select is consequently central in many approaches to the design and development of user interfaces.

Discussions of how to measure the quality of computer systems have gone on for several decades, first under the heading of ergonomics (Shackel, 1959) and ease-of-use (Miller, 1971; Bennett, 1972), and later under the heading usability (Bennett, 1979; Shackel, 1981). However, recently discussions recur on which measures of usability are suitable and on how to understand the relation between different measures of usability. These discussions of usability are in part fueled by concerns on the limitations of commonly employed usability measures. Take as an example Dillon (2001), who has argued that users, designers, and owners of a system may not equally weight the importance of a usability measure such as time. Thus, the importance of time as a measure of usability may be overestimated. Another example comes from Hassenzhal et al. (2000). They argue that commonly employed usability measures ignore what they call hedonic quality, that is "quality dimensions with no obvious relation to the task the user wants to accomplish with the systems, such as originality, innovativeness, beauty, etc." (Hassenzahl et al., 2000, p. 202).

Another source of the discussion of usability is the challenge to established usability techniques and measures from new contexts of use such as home technology (Monk, 2002), ubiquitous computing (Mankoff et al., 2003), and technology supporting learning (Soloway et al., 1994). These contexts of use, it is often argued, require new measures of usability to adequately capture what is considered important in the particular context. For example, Monk (2002) has argued that when we look at technologies for homes, activities may not be undertaken to reach a certain goal but rather to get certain experiences. Consequently, non-traditional usability measures and evaluation techniques have to be applied.

Finally, proposals for new measures of usability are continuously emerging. The HCI literature now contains discussions of, for example, fun (Carroll and Thomas, 1988), aesthetics (Tractinsky, 1997), apparent usability (Kurosu and Kashimura, 1995), sociability (Preece, 2000), and flow (Hoffman and Novak, 1996). These proposals all seem to suggest that common conceptions of how to measure usability may need revisiting.

On this background, the present paper reviews current practice in measuring usability, giving an empirical basis for discussing how to measure usability. From the review we outline challenges regarding on the one hand how to conduct usability studies, especially concerning which usability measures to select, and on the other hand what to research so as to improve the validity and reliability of usability measures. The former challenge is especially relevant to researchers and practitioners who conduct usability studies; the latter is relevant mainly to usability researchers. In addition, we intend the paper to serve for practitioners and researchers alike as a catalogue of the variety in which usability may be measured.

Next, Section 2 presents the method used for reviewing the usability measures employed in a selection of studies from high-quality HCI journals and conferences. Section 3 summarizes and discusses the measures used. Section 4 discusses the challenges identified and gives some suggestions on how to better measure usability.

2. Method of review

To understand better the current practice in measuring usability, this section describes the method for the review of usability measures employed in 180 studies from the HCI research literature.

The review is based on a broad conception of usability, similar to that of ISO 9241, part 11 (1998) and Bevan (1995). This conception means that we include studies in the review that measure aspects of quality-in-use that fall under our definition of usability, though not all studies use the term usability to denote the measures employed. Insisting on a mention of the term usability would lead to a small sample, which is why we use a broad definition of usability.

2.1. Selection of studies

The studies considered as candidates for inclusion in the review were chosen from the last two, full volumes of a selection of HCI proceedings and journals available at the time of writing, see Table 1. The proceedings and journals chosen are among the core forums for publishing HCI research. A mix of studies from both journal and proceedings was considered as experimenting with measures of usability are most likely to appear in proceedings,

Table 1
Candidate studies and studies selected for inclusion in the review

Journals/proceedings	Years	Candidate studies		Studies selected	
		N	%	N	%[a]
ACM Transactions on Human–Computer Interaction	2000–2001	34	6	10	6
Behaviour & Information Technology	2000–2001	86	15	25	14
Human–Computer Interaction	2000–2001	14	2	1	1
International Journal of Human–Computer Studies	2000–2001	163	28	32	18
ACM CHI Conference	2001–2002	135	23	56	31
IFIP INTERACT Conference	1999, 2001[b]	155	26	56	31
Total		587		180	

[a]percentages do not add up because of rounding errors.
[b]INTERACT is only held bi-annually.

while journals might be more representative of carefully conducted and thoroughly reviewed studies. We include only full-length and original research papers, excluding review papers and old papers in reprint. In some cases, a paper contained two or more experiments or comparisons, each differing in their choice of usability measures. In these cases, we treat those experiments or comparisons as individual studies. In this way, a total of 587 candidate studies were identified.

For a study to be included in the review, three criteria should be met. First, the study should report quantified data on usability by describing a measure of usability in either the section describing the method of the study or in the section reporting the results. We exclude studies that only present experiences with user interfaces. As an example, a paper on a wearable audio device (Sawnhey and Schmandt, 2000) reported use experiences from one of the authors for a period of more than 1 year and from two users interacting with the device for 3 days. The paper reports various practical and conceptual problems experienced by these users, yet it contains no attempt to quantify usability and is thus excluded from the review. Likewise, informal usability tests reporting only selected comments from users were also excluded.

Second, we include only candidate studies that evaluate the quality of interaction between human users and user interfaces in a broad sense (paper prototypes, information appliances, software, etc.). Studies that focus on comparing the predictions of cognitive models to the behavior of actual users are thus excluded, as they reason about the fit between models and behavior, not about the quality of interaction. A number of studies report qualitative data on algorithmic effectiveness or effectiveness of the system part of speech recognition systems; they are also excluded.

Third, to be included a candidate study should describe comparisons that are relational (Rosenthal and Rosnow, 1991, p. 10 ff.), that is it should use usability measures to characterize differences in the interaction with user interfaces. Such differences could be, for example, between computer systems, use situations, user groups, organizations, task types, previous versions of the system, over time, or anything else where the focus is on comparing the quality-in-use. We employ this criterion to exclude studies that are not focused on using usability to compare interfaces, but for example concern the demographics of Internet users.

For each study, we collected information about the usability measures employed. We were liberal in including measures as a measure of usability, but did not include system-related measures (response times) or personality measures (e.g., intelligence). In addition to the usability measures used, we recorded for each study (1) if a special instrument was used for any of the usability measures employed; (2) if validation of the measures employed were made; (3) the duration of interaction with the user interfaces; and (4) if the study included data on the correlation between usability measures, such as correlation coefficients or results of factor analysis. If the duration of interacting with interfaces was not explicitly mentioned, we estimated it based on information on task completion time, where available.

To assess the reliability of inclusion of a candidate study into the sample and the identification of usability measures in each study in the sample, we had two independent raters classify a random sample of 20 studies, 10 included in the review and 10 that were not. Cohen's kappa for agreement on inclusion in the review was .8 and .9 for the two raters; average kappas for points (1)–(4) above were .9, .8, .8, and .9. Overall, this suggests an excellent agreement among raters (Fleiss, 1981).

The selection of studies described above gave a sample of 180 studies (see Appendix A for a list of references to the studies; 14 of the references include two studies). Note that the sample is somewhat biased towards conference proceedings, both in terms of studies considered as candidates for the review (49%) and in terms of studies selected for the review (62%). However, historically and in the opinion of many HCI researchers, papers from the CHI conference (and to a some extent also the Interact conference) serve as exemplary cases of research in HCI. Therefore we consider the sample representative of how HCI researchers measure usability.

2.2. Classification of usability measures

To gain an overview of the measures employed, we classified them into the three groups effectiveness, efficiency, and satisfaction of the ISO 9241 standard for usability (ISO, 1998). The ISO standard defines usability as the "[e]xtent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction". Further, effectiveness is the "[a]ccuracy and completeness with which users achieve specified goals"; efficiency is the "[r]esources expended in relation to the accuracy and completeness with which users achieve goals"; and satisfaction is the "[f]reedom from discomfort, and positive attitudes towards the user of the product" (ISO, 1998, all p. 2). The ISO standard was chosen as a basis for classification because its three groups of measures are widely accepted to concern distinct measures and because it seemed instrumental in establishing a first overview of the measures used. Section 4.7 will discuss in more detail this choice and some of the limitations of the ISO standard.

In each of these groups we further classified measures into subgroups. These subgroups were based in part on usability measures mentioned in prominent text books (Nielsen, 1993; Dix et al., 1993; Shneiderman, 1998), in a book on behavioral science (Meister, 1985), and in well-known discussions of usability measures (Whiteside et al., 1988; Sweeney et al., 1993); and in part on the similarities found in the usability measures used by the studies reviewed. It is not the intent of this paper to create a standard classification, merely to group usability measures so as to make similarities and differences in current practice clear. Subgroups containing only one or two instances of measures were merged into an "other" subgroup. As will be evident from the review, the subgroups are not wholly exclusive.

To assess the reliability of the classification into groups, we had the independent raters mentioned above classify measures into groups and subgroups. Average agreement for this classification was 89.7%, suggesting an excellent agreement among raters.

3. Current practice in measuring usability

Below we describe the usability measures employed in the 180 studies by discussing in turn effectiveness, efficiency, and satisfaction measures. The purpose of this section is to put forward data on measures used in the studies and to give some immediate comments. Section 4 contains a broader discussion on the challenges posed to usability studies and research by the current practice in measuring usability.

3.1. Measures of effectiveness

Table 2 summarizes the measures of effectiveness employed in the studies reviewed. The row labelled *binary task completion* refers to measures of whether users complete tasks or not. Binary task completion includes measures of the number of correct tasks, the number of tasks where users failed to finish within a set time (e.g., Westerman et al., 2001), or the number of tasks where users gave up (e.g., Dumais et al., 2001).

Accuracy measures quantify the number of errors users make either during the process of completing tasks or in the solution to the tasks. For example, Marshall et al. (2001) counted the number of errors in data entry, and Nyberg et al. (2001) had the number of hints given by the experimenter subtract from accuracy. In addition, Table 2 shows two subgroups of accuracy measures. Spatial accuracy is typically used in studies of input devices and is measured as the distance to a target from the position indicated by the user; precision is typically used in studies of information retrieval systems and indicates the ratio of the number of correct documents retrieved to the total number of documents retrieved.

Recall refers to measures of how much information users can recall after having used the interface. Typically, the information recalled is parts of the content of the interface. Bayles (2002) for example, investigated users' recall of banner ads and if animation influences recall. She measures recall as users' ability to reconstruct web pages seen and their ability to differentiate between banner ads and distracter ads.

Completeness refers to measures of the extent to which tasks are solved. Such measures are orthogonal to accuracy measures as they capture effectiveness in situations where it does not make sense to say that users made errors in completing tasks, but only that users reached solutions of different completeness. Examples of such measures include the number of secondary tasks solved (McFarlane, 1999) and the proportion of relevant documents found in information retrieval tasks (Cribbin and Chen, 2001).

Quality of outcome is a more extensive attempt to measure the outcome of tasks, for example the quality of a work product or of learning. The criteria used for differentiating between measures of quality and of accuracy are that the latter measure of usability can be measured unambiguously and often automatically; measures of quality are more complex to obtain and aim to capture richer aspects of the outcome of interaction. Measures of quality include measures of understanding, for example tests of what has been learned from an instructional interface. Corbett and Anderson (2001) assessed the effectiveness of computer-based tutors by paper-and-pencil tests on programming abilities. Hornbæk and Frøkjær (2001) used a five-point scale to grade the quality of essays written by users of three interfaces for electronic documents. In addition, Hornbæk and Frøkjær used subjects' own grading of the essays written as a measure of the subjective perception of the outcome of the interaction. In a similar vein, Gong and Lai (2001) had two judges code on a 0–3 rating scale the ease and involvement with which users complete tasks.

Table 2
Measures of effectiveness

Measure	N	%	Explanation	Examples from the studies reviewed
Binary task completion	24	13	Number or percentage of tasks that users successfully complete	Number of correct tasks (e.g., correctly selected icons); number of tasks that users do not complete in the allotted time; number of tasks where users give up; number of incomplete tasks
Accuracy	55	31	The accuracy with which users complete tasks, that is some quantification of error	
Error rates	46	26	Errors made by the user during the process of completing a task or in the solution to the task	Number of errors on the way to task completion (e.g., syntax errors in searching); number of errors in a series of tasks; number of correctly done subtasks; percentage correct solutions; number of hints needed to complete a task; tasks-to-criterion (e.g., number of attempts to achieve two correct tasks)
Spatial accuracy	7	4	Users' accuracy in pointing to or manipulating user interface objects	Distance from target (e.g., mm from correct icon); orientation error in radians when rotating a virtual object
Precision	3	2	The ratio between correct information and total amount of retrieved information	Precision in information retrieval systems
Recall	11	6	Users' ability to recall information from the interface	Recall of information presented in the interface (e.g., recall of banner adds); recall of features of the interface (e.g., memory for button location)
Completeness	11	6	The extent or completeness of users' solutions to tasks	Number or percentage of secondary tasks done; number of relevant documents identified; payoff in a virtual stock broker task
Quality of outcome	28	16	Measures of the quality of the outcome of the interaction	
Understanding	18	10	Understanding or learning of information in the interface	Multiple-choice tests on information presented in the interface; pre-test to post-test difference in understanding (e.g., of programming languages); standardized tests of learning (e.g., Graduate Record Examinations)
Experts' assessment	8	4	Experts' assessment of outcomes of the interaction	Expert grading of work products
Users' assessment	3	2	Users' assessment of the outcome of interaction	Users grading of the perceived quality of work products; agreement among users on relevant documents
Other	6	3	Other measures of effectiveness	Users' ability to predict functioning of interface after use; dispersion of balance on a platform when using virtual environments; users' health information
Any effectiveness measure	117	65	Studies that contain any of the above measures	
Effectiveness controlled	23	13	Studies that consider only correctly completed tasks	Analyse only correct tasks; use very simple tasks (e.g., tasks take less than 2s); prevent users from continuing before task is correctly solved; task cannot be wrong; other measures ensures that only correct tasks are considered (e.g., input rate measured as corrected words per minute)
No effectiveness measure	40	22	Studies that do not report any measure of effectiveness, nor control effectiveness	

Note: Counts and percentages show number of studies using *any* measure falling under the description.

Other measures of effectiveness include users' ability to predict the functioning of the interface and changes in users' health.

The row labelled *effectiveness controlled* refers to studies that only consider correctly completed tasks (e.g., by allowing the user only to proceed only with correct answers) or that use other measures that indirectly controls effectiveness (e.g., corrected words entered per minute).

Four comments on the measures of effectiveness used are pertinent. First, 22% of the studies reviewed do not report *any* measure of effectiveness nor do these studies control effectiveness. Frøkjær et al. (2000) analysed a sample of 19 studies that use complex tasks and found two studies that did not include any measures of effectiveness. They described how such studies may reach unreliable overall conclusions about usability because they do not measure effectiveness, for example by claiming that one system is superior to another based only on counts of the resources spent in communication (an efficiency measure). In addition, Frøkjær et al. argued that the HCI community might not succeed in trying to make better computing systems without employing measures of effectiveness in all studies. Table 2 indicates that these problems can be found broadly in HCI research.

Second, Table 2 shows that measures of the quality of the outcome of the interaction are used in only 16% of the studies. For example, experts' assessment of work products seems a solid method for judging the outcome of interaction with computers and has been used in a variety of fields as an indicator of the quality of work products, for example with respect to creativity (Amabile, 1996). Yet, in our sample only 4% of the studies use such measures.

Third, as mentioned in the introduction a recurrent theme in the discussion of usability is whether new kinds of devices and use contexts require new measures of usability. Especially, it has been argued that the notion of task underlying any effectiveness measure will not work in emerging focuses for HCI, such as home technology. In the present sample of studies, however, a range of measures is used that certainly do not rely on a confined notion of task, yet can easily be placed in the ISO classification. For example, the measures reviewed include measures of the quality of life (Farnham et al., 2002) and the gain of virtual money in stock broker support systems (Bos et al., 2002). Thus, the framework of evaluation may not be the problem; rather what seems challenged is researchers' ingenuity in devising new measures of effectiveness.

Fourth, a number of studies combine usability measures into a single measure, report the combined values, and make statistical tests on the combinations. A combination here means a function that takes as arguments two or more individual usability measures; most often this includes an effectiveness measure. As an illustration, consider a study of a voicemail interface (Whittaker et al., 2002). In this study, three independent judges rated the tasks; in addition, task completion times were collected. The authors observed that:

> Users differed greatly in their thoroughness: some spent large amounts of time trying to optimize solutions, whereas others were satisfied with quick, but approximate, solutions. As in our previous research [19], to control for this, we used a normalized measure of success, namely (quality of solution)/(time to solution) (Whittaker et al., 2002, p. 279).

They then proceeded to present means of the normalized measure of success, and to perform analysis of variance using this measure. In addition to the reason cited above for using this normalized measure of success, there are other reasons for using combined measures such as simplifying data analysis and decreasing the risk of inflating Type I error by doing multiple tests. Indeed, Whittaker et al. used the normalized measure of success for examining the general hypothesis that their interface would outperform an existing interface. However, studies using combinations of measures share some of the problems discussed for studies without effectiveness measures. In the study of Whittaker et al. (2002), how do we know that differences between systems are not only due to differences in time to solution? Implicitly, their normalized measure contained a weighting of the importance of time-to-solution differences versus differences in the quality of solutions. Such weighting may be hard to do a priori, but will depend on the particular scales of measurement if no weighting is chosen. In summary, we find that the combined measures do not lead to clarity in analysing and reporting the results; moreover, they seem to run the risk of finding differences in overall performance, where there exists only a difference in, say, the time used to complete a task.

Another example of a combined measure is throughput based on the effective width of targets, used in several studies of input techniques included in this review (e.g., MacKenzie et al., 2001). It has been argued that also this combination of measures may hide underlying patterns (Zhai, 2004).

3.2. Measures of efficiency

Table 3 summarizes the measures of efficiency used in the studies reviewed. The row *time* refers to measures of how long users take to complete tasks with the interface. Fifty-seven percent of the studies measure time as task completion time. Some studies, however, measures time for different modes of the interaction, such as the time taken for parts of a task, the time spent in the help function, or the time used in different parts of the interface. Examples of such measures are mean time between actions (Golightly et al., 1999), the time spent in dialog boxes (Tyfa and Howes, 2000), or the time dwelled in certain parts of the display (Burns, 2000). Other studies measure the time elapsed before a certain event. An example of such measures is the time to the first key press (Mitsopoulos

Table 3
Measures of efficiency

Measure	N	%	Explanation	Examples from the studies reviewed
Time	113	63	The duration of tasks or parts of tasks	
Task completion time	103	57	The time users take to complete a task	Time to complete a task
Time in mode	26	14	The time users spend in a particular mode of interaction, e.g. on part of a task or in part of the interface	Time on parts of the task; duration of use of particular functions or interfaces; time per action; mean duration of pauses between actions; time used in help; time used on searching versus time used for pointing
Time until event	10	6	Time elapsed until users employ a specific feature or perform a particular action	Time before a secondary task is solved; time to react to a warning
Input rate	12	7	Rate of input by the user, for example using mouse or keyboard	Words per minute entered; throughput; correctly entered words per minute
Mental effort	9	5	The users' mental effort when using the interface	NASA's Task Load Index questionnaire (Hart and Staveland 1988); task difficulty rated by experts; physiological measures of effort; ratings of mental effort by users
Usage patterns	44	24	Measures of how users make use of the interface to solve tasks	
Use frequency	23	13	The frequency of function use or actions	Number of keystrokes; number of mouse clicks; number of functions used; number of interface actions; amount of mouse activity; number of times help is consulted
Information accessed	13	7	The amount of information that users access or employ	Number of web pages visited; number of buttons looked at
Deviation from optimal solution	13	7	The ratio between actual behavior and an optimal method of solution	Theoretical optimal minus actual distance traveled in virtual environments; number of extra actions taken to complete a task
Communication effort	5	3	Resources expended in communication processes	Number of speakers' turns; number of words spoken; number of interruptions, amount of grounding questions asked
Learning	5	3	Users' learning of the interface	Changes in task completion time over sessions
Other	17	9		Reading speed; measures of lostness; amounts of functions used; number of times users switch between parts of the interface
Any efficiency measure	134	74	Any of the above measures	
Time controlled	12	7	Fixed time for task completion	Fixed time for tasks; very simple tasks
No efficiency measure	34	19	The study does not report any measure of efficiency, nor is efficiency controlled	

Note: Percentages and counts show number of studies using *any* measure falling under the description.

and Edwards, 1999) or the time it takes to get to the first relevant node in a hypertext (Cribbin and Chen, 2001).

Input rate is measured in a number of studies, typically in the form of text entry speed (words per minute, corrected words per minute) or throughput. For example, Isokoski and Käki (2002) measured the average number of correctly entered digits per second for two input methods.

Mental effort concerns the mental resources users spend on interaction. Five of the studies reviewed measure mental effort using NASA's Task Load Index (Hart and Staveland, 1988). Four studies use other measures of mental effort, including measures of heart rate variability (Izsó and Láng, 2000) and subjective time estimation (Tractinsky and Meyer, 2001).

Usage patterns include measures of how the interface is used. The rationale behind these measures is that the usage of the interface is indicative of the resources users expend when trying to solve a task. One kind of usage pattern measure simply focuses on the number of times a certain action has been performed. For example, Marshall et al. (2001) counted the number of keystrokes used for completing alpha numerical data entry; Drucker et al. (2002) counted the number of mouse clicks needed to complete video browsing tasks. Another kind of usage pattern aims to measure how much information users access when solving tasks, for example the number of objects visited in a virtual space (Westerman and Cribbin, 2000). A third kind of usage pattern consists of measures of the deviation from the optimal solution, using the relation between the actual behavior of the user and the optimal efficient solution to a task. These measures include the ratio of actual distance traveled to shortest distance in traveling information spaces (Tan et al., 2001) and the number of target re-entries in input studies, that is the number of times the mouse pointer enters a target region, leaves, and then re-enters the target area (MacKenzie et al., 2001).

Communication effort refers to measures of the resources users expend in communication, typically employed in studies of groupware. Examples include number of turns in conversation (Matarazzo and Sellen, 2000) and the number of grounding questions used in cooperation between users in a shared browsing environment (Farnham et al., 2000).

Learning measures use changes in efficiency as an indicator of learning, for example in the time used for completing tasks. Note that these measures are not necessarily related to measures of effectiveness that address learning, because the latter concern learning facilitated by the content of the interface. We also do not consider studies that briefly look at an ordering effect of interfaces in experimental studies to be measuring learning. Examples of learning measures are found in the study of Isokoski and Käki (2002), that investigated how users becomes faster in text input over time.

Other measures include three studies measuring reading rate in words per minute and a study of expert assessment of task completion ease.

Time controlled refers to studies where users are given a fixed amount of time to complete their tasks. The fixed amount of time can be enforced by a particular procedure (stopping subjects) or by the nature of the task (fixed time for looking at an interface). For example, Tractinsky and Meyer (2001) studied the influence of menu layout on time estimation. In their study, all tasks had the same duration of 12 s so as to establish a uniform base from which to compare time estimations.

Five comments on the measures of efficiency used seem relevant. First, some of the efficiency measures are obviously related to the quality of interactive computer systems, because they quantify resources (e.g., time or mental effort) that are relevant in many contexts for many users. In our view, the efficiency measures summarized under the phrase usage patterns do not show the same obvious connection to quality-in-use.

Take as an example the study by Woodruff et al. (2001). They evaluated three ways of making summaries of web pages. In a user study, they compared the usability of these summaries by measuring task completion time, number of web pages visited, and users' preference (in the description of the experiment, it is unclear if effectiveness was measured or controlled). In the paper, Woodruff et al. made a comprehensive analysis of the task completion times with the three interfaces. They also analysed the number of web pages visited, possibly wanting to support their data on task completion times with an alternative measure. However, in the paper Woodruff et al. seem to treat the number of web pages visited as a usability measure, that is as something indicative of quality-in-use. For example, they referred to that measure under the umbrella term 'performance', e.g., in beginning their discussion:

> As one might expect, the relative performance of text summaries, plain thumbnails, and enhanced thumbnails depends greatly upon the question category. For the Picture question, the text summaries required more search time and more pages visited than either type of thumbnail (Woodruff et al., 2001, p. 204).

The question raised is what does the number of web pages visited signify as an indicator of usability?

We see the number of web pages visited is an indirect and interface-oriented usability indicator. It is indirect because the number of web pages visited seems irrelevant, if it is not reflected in task quality, task completion time, or subjective satisfaction. It is interface-oriented because it is entirely possible that users' perception of navigation difficulty is unrelated to the number of web pages visited. Such indicators are useful mainly as descriptions of the use of the system, not as indicators of the quality of interactive systems. Woodruff et al. thus seem on the verge of treating usage patterns as an indicator of how well an interface support users in solving their tasks.

Another example of a measure of usage patterns with no straight-forward interpretation comes from a study by

Drucker et al. (2002) of how users navigate video. In their study, the number of clicks was counted, in addition to measures of task completion time, accuracy and subjective satisfaction. Here, it is unclear what number of clicks has to do with usability. The authors reported: "We assessed differences on 3 performance measures: time to complete the tasks, the number of clicks to complete tasks and accuracy in performance" (Drucker et al., 2002, p. 223) and seemed to consider the number of clicks as a measure suitable for distinguishing between good and poor interfaces. However, why should we pay attention to number of clicks if subjective satisfaction, effectiveness, and time measures show other patterns? Note that this is not to say that the number of clicks has no use in design considerations (see for example Card et al., 1983 or Raskin, 2000) or as a description of usage that may help HCI researchers understand how the interface is used. Perhaps this is what Ducker et al. intended, as one of their hypotheses reads "User performance (measured by speed and accuracy with which they complete the tasks)" (p. 221); this careful choice of indicators of user performance, however, is lost in the quote above and in their abstract.

A second comment on the studies reviewed pertains to the measurement of time. A surprising pattern apparent from Table 3 is that while objective task completion time is measured by 57% of the studies, little attention is paid to users' experience of time. A number of studies measure subjectively experienced workload, and, as we shall see when discussing satisfaction measures, a number of studies also measure subjects' experience of task difficulty. However, in this sample of 180 studies, only one study measures directly subjective experience of time, namely Tractinsky and Meyer (2001).

Third, the reviewed studies differ in how task completion times, and efficiency measures in general, are reasoned about. In the ISO definition of usability and in most of the studies reviewed, time is considered a resource of which successful interfaces minimize consumption. However, in a handful of studies higher task completion times are considered as indicators of motivation, reflection, and engagement. As an example, consider the study of Inkpen (2001). She compared a point-and-click to a drag-and-drop version of a puzzle solving game. Participants were asked to solve as many puzzles as possible in 30 min and told that they could stop playing when they wanted. Motivation was measured as whether or not the girls played for the full 30 min; motivation was assumed to be higher if the girls played for the full 30 min. In comparing to other studies, there are important differences in task types and instructions for users. Yet, imagine for a moment the sweeping implications of Inkpen's reasoning about time on the understanding of the results from at least some of the studies discussing time as a limited resource.

Fourth, a striking pattern among the studies reviewed is that so few measures concern learning of the interface. Only five studies measure changes in efficiency over time, all investigating input techniques. We return to these studies in Section 4.

Fifth, in the studies reviewed, the median time of working with the user interfaces evaluated was 30 min (the average is 76.7 min, standard deviation is 239.4 min); the large majority of studies had users complete their tasks in just one session.

3.3. Measures of satisfaction

Tables 4 and 5 summarize the measures of satisfaction used in the studies reviewed. Many studies do not give details on the questionnaires used for assessing satisfaction; only the construct they were intended to measure is referred to. Robertson et al. (2002), for example, mention only that they measured average satisfaction rating, but do not write the actual questions they asked users. Below, we group such studies into subgroups according to the construct mentioned. For studies that mention both the construct and the actual questions, and report validation or reliability efforts, we group the satisfaction measures according to the construct measured. The study by Isbister and Nass (2000) construct from factor analysis an index of fun, consisting of adjective items based on the terms enjoyable, exciting, fun, and satisfying; it is classified under Fun. For studies that mention questions for users and that do not fall in the above two groups, we split up the questions into different subgroups—questions that may have been reported together by the authors of the study.

The row labeled *standard questionnaires* refers to studies that use standard questionnaires for measuring satisfaction or build directly upon previous work for questions on overall user satisfaction. Among the studies reviewed, the questionnaire for user satisfaction, QUIS (Chin et al., 1988; Shneiderman, 1998, pp. 134–143), is used in 4% of the studies.

Preference measures capture which interface users prefer using. Usually, preference is measured by forcing users to choose which interface they prefer, or in the case of more than two interfaces, to rank the interfaces according to preference. Preference measures can be obtained for instance by asking users "which system did they prefer for the targeting task" (Gutwin, 2002) or by asking users to rank-order five interfaces from 1, like the most, to 5, like the least (Rui et al., 2001). Another approach is to let users rate each interface on a scale, prompting for their preferred or most liked interface—in Wang et al. (2001, p. 315) "subjects scored their preference about markup methods in 5 levels (5—like very much, 4— like, 3—neither like nor dislike, 2—slightly dislike, 1—dislike)."

Ease-of-use refers to measures of general satisfaction with the interface, intended to measure the same construct as the standard questionnaires. Such measures are obtained by having users respond to questions that typically use phrases such as satisfaction/satisfied/user satisfaction, for example "rating scales 0 = very low; 6 = very high […] were you satisfied with the MDS used" (Matarazzo and

Sellen, 2000, p. 342); such as easy to use/ease of use/ease, for example "7-point Likert scale from 'strongly agree' to 'strongly disagree' […] The method was easy to use" (Marshall et al., 2001, pp. 175, 186); or such as useful/usefulness, for example "Usefulness of the interaction was an index used to characterize the interaction: helpful and useful" (Isbister and Naas, 2000).

Some of these questions ask users to consider specific functions of the interface or to answer questions on ease-of-use only relevant in the particular context of use. Examples of such questions include the following, used in a study comparing search engines and a novel interface for searching: "How satisfied were you with the use of the refine function provided by the search engine (alta vista)?" and "How satisfied do you feel with the use of the keyword comparison?" (Fang and Salvendy, 2000, p. 924). Most of these measures are based on questions using the phrases preference, liked best, easy, useful, and compares well to.

A few studies measures ease-of-use *before* users interact with the interface. For example, Cockburn and McKenzie (2001) asked users to rate on a five-point Likert scale "I will be able to quickly find pages" (p. 437) after users have spent time organizing web pages but before they have tried to find any pages; other studies try to measure satisfaction *during* users' interaction with interfaces. As an example, Izsó and Láng (2000) used heart rate variability during interaction to predict perception of task difficulty.

Specific attitudes include measures aimed at capturing specific attitudes toward or perceptions of the interface; Table 5 summarizes the measures employed. The most common attitudes mentioned in the studies reviewed are *liking* for example rated on a scale from 1 (disagree) to a 5 (agree) "I like the software I used today" (Czerwinski et al., 1999, p. 169); *fun* rated on a 1 to 8 scale: "I thought the interface was fun to use" (Drucker et al., 2002, p. 223); *annoyance* "Participants were asked to rate each cue for perceived annoyance. […] 5 = very annoying, 1 = very pleasant" (Pacey and MacGregor, 2001, p. 174); and feeling of control, "it's easy to make the software do exactly what I want" (McGrenere, et al., 2002, p. 168). Note that we include measures of liking under Specific attitudes, though the questions used to measure liking seem very similar to questions used in preference measures that use rating. However, in the study of Rui et al. (2001) there appeared to be a difference, in that the five interfaces are ranked differently depending on which of the following questions are used: (1) "Rank order of the interface (1 = like the most, 5 = like the least)" and (2) "Ratings: I liked this interface (1 = strongly disagree, 5 = strongly agree)", (Rui et al., 2001, p. 455). Thus, it is unclear if these kinds of liking and preference questions in general may be considered similar.

A whole range of measures is concerned with users' attitudes and perceptions of phenomena other than of the interface itself, for example of other persons, of the content of the interface, and of the process and outcome of interaction. *Attitudes toward other persons* refer to measures used for assessing users' attitudes towards communication and collaboration partners. Such measurers typically aim at capturing the feeling of presence, trust, common ground, and ease of communication. One study had subjects complete a social anxiety test in addition to answering questions on their sense of being together (Basdogan et al., 2000). In some studies, measures of attitudes and perceptions of other persons are applied to user interfaces, for example to evaluate agents. Brickmore and Cassell (2001), for example, had users complete a standard questionnaire on trust (taken from Wheeless and Grotz, 1977) after interacting with a conversational agent.

Other measures ask users about their attitudes towards the *content of the interface*. Such questions could be about the quality of the information, the interest subjects took in

Table 4
Measures of satisfaction

Measure	N	%	Explanation	Examples from the studies reviewed
Standard questionnaires	12	7	Measure satisfaction by using a standardized questionnaire or by building directly on previous work	Questionnaire for user satisfaction, QUIS (Chin et al., 1988); questions from Davis (1989)
Preference	39	22	Measures satisfaction as the interface users prefer using	
Rank preferred interface	29	16	Users chose or rank interfaces according to preference	"Which interface did you prefer"; "Indicate one preferred tool"
Rate preference for interfaces	5	3	Users rate the preference of each interface	Rate preference on a 1 to 10 scale; "preference about markup methods in 5 levels"
Behavior in interaction	5	3	The preferred interface is indicated by users' behavior in interaction	Enable users to continually chose an interface to perform tasks with; observe which interface facilities users chose to use
Satisfaction with the interface	65	36	Users satisfaction with or attitudes towards the interface	
Ease-of-use	37	21	Broad measures of users' overall satisfaction or attitudes towards the interface or user experience	"This software is satisfying to use"; "satisfaction with the interaction", "this interface was easy to use"; "overall I think this

Table 4 (*continued*)

Measure	N	%	Explanation	Examples from the studies reviewed
				is a good system''; ''overall quality of the user interface''; ''this site compared favorably to others I have used''
Context-dependent questions	36	20	Users' satisfaction with specific features or circumstances in the specific context of use	''I would use this site if I needed more information about this organization''; ''navigation through the menus and toolbars is easy to do''; ''I could skip commercials easily with this interface''; ''it is clear how to speak to the system''
Before use	4	2	Measures of satisfaction with the interface obtained before users have interacted with the interface	''I will be able to quickly find pages''
During use	3	2	Measures of satisfaction obtained while users solve tasks with the interface	Heart period variability; reflex responses; quantifications of negative comments during use; counting of users getting nausea
Specific attitudes	39	22		
Users' attitudes and perceptions	44	24	Users' attitudes towards and perceptions of phenomena other than the interface	
Attitudes towards other persons	19	11	Measures of the relation to other persons or to interfaces, considered as persons	''I felt connected to the other persons in my group''; feelings of trust (Wheeless and Grotz, 1977); ''impression of conversion''; social richness; sense of being together (Slater et al., 2000); ''which character did you feel familiar with?''
Attitudes towards content	8	4	Attitudes towards the content of the interface when content can be distinguished from the interface	''the information was of high quality''; ''How appealing was the subject matter''; ''novelty of the articles read''
Perception of outcomes	12	7	Users' perception of the outcome of the interaction	''How do you judge the quality of the task outcome?''; users' sense of success; assessment of own performance
Perception of interaction	17	9	Measures users' perception of the interaction	''With which interface did you think you were faster?''; users' perception of task difficulty
Other	25	14	Other measures of satisfaction	''Pleasantly surprising''; ''easy to make mistakes''; ''meaningfulness''; ''job satisfaction''; ''I felt the method was reliable''; ''naturalness''; ''embarrassment during task''; Short Symptoms Checklist; feelings of presence; feedback from the interface; ''the display is cluttered''
Any satisfaction measure	112	62	Any of the above measures	
No satisfaction measure	68	38	None of the above measures	

Note: Percentages and counts show number of studies using *any* measure falling under the description. The examples in quotes are questionnaire items, most often answered on a five or seven-point Likert scale.

the information, or the organization of the information. As an example, Karat et al. (2001, p. 460) asked users ''How appealing was the subject matter of the multimedia experience to you? (1 = Not appealing at all, 7 = Very appealing)''.

Perception of outcomes refers to users' rating of their perception of the outcomes of the interaction. This is measured as answers to questions on users' confidence in the solution to tasks, as users' perception of comprehension, as users' perception of learning, or as users' assessment of their own performance. As an example that includes several measures of the perception of outcomes consider a study of the understanding of presentations (LeeTiernan and Grudin, 2001). In that study, users were asked to rate on a seven-point scale ''Overall I am satisfied with my work product'', ''I am convinced that my final arguments are best'' and ''I learned as much as possible from this lecture'' (LeeTiernan and Grudin, 2001, pp. 476–477).

Perception of interaction refers to users' rating of their perception of the process of interaction. This most often regards users' perception of task complexity and of task completion times. As a representative study consider Corbett and Anderson (2001). In this study, users of a

Table 5
Measures of specific attitudes towards the interface

Measure	N	%	Explanation	Examples from the studies reviewed
Attitudes towards interface	39	22	Questions given to users aiming to uncover specific attitudes towards the interface	
Annoyance	7	4	Measures of annoyance, frustration, distraction and irritation	"I thought this interface was frustrating to use"; "I felt flustered when using this method"; user experience going from very comfortable to very frustrated
Anxiety	3	2	Users' anxiety when using the interface	Self-evaluation state anxiety form (Speilberger, 1983)
Complexity	3	2	Users' perception of interface complexity	Complexity versus order of interfaces
Control	7	4	Users' sense of control and attitude towards the level of interactivity.	"It's easy to make the software do exactly what I want"; "extent to which the system enables the subjects to actively interact with it"
Engagement	4	2	Users' experience of engagement, involvement and motivation	"How engaging was the multimedia experience for you?"; User experience of enthusiasm and motivation
Flexibility	3	2	Users' perception of flexibility in the interface	Flexibility of the interface
Fun	14	8	Users' feeling of fun, entertainment, and enjoyment	"It was enjoyable to use"; "I thought this interface was fun to use"; "how entertaining was the multimedia experience for you"
Intuitive	3	2	Users' perception of the intuitiveness of the interface	"Layout is intuitive"; "intuitive to use"
Learnability	5	3	Users' attitude toward how easy it is to learn to use the interface	"I was able to learn how to use all that is offered in this software", "I found this interface easy to learn"
Liking	15	8	Users' liking of the interfaces	"I liked this interface?", "I liked the software I used today"; rate the interface between love and hate
Physical discomfort	3	2	Users' experience of physical discomfort in using the interface	"Eyes become sore", "upper body discomfort"; total muscular discomfort; physically tiring
Want to use again	3	2	Users' attitude towards using the interface again	"Would be happy to use again"; "I would like to have the interface available for my use all the time"

Note: The examples in quotes are questionnaire items, most often answered on a five or seven-point Likert scale.

tutoring system aimed at supporting programming answered on a seven-point scale two questions: "How difficult were the exercises" and "Did the tutor help you finish more quickly" (Corbett and Anderson, 2001, p. 251).

Other measures include measures of beauty, how cluttered subjects find a display, and a measure of users' embarrassment.

Four comments on the satisfaction measures used are relevant. First, the measurement of satisfaction seems in a state of disarray. A host of adjectives and adverbs are used, few studies build upon previous work, and many studies report no or insufficient work on the validity and reliability of the instruments used for obtaining satisfaction measures. The diversity of words used in five-point or seven-point semantic differentials or Likert-type rating scales is simply astonishing; they are typically used in questions similar to "The system is …" or "I feel …" and include:

> Accessible, adequate, annoying, anxiety, appealing, boring, clear, cluttered, comfortable, competent, comprehensible, conclusive, confident, conflict, confusing, connected, convenient, desirable, difficult, dislikable, dissatisfied, distracting, easy, effective, efficient, embarrassed, emotional, engaging, enjoyable, entertaining, enthusiasm, excellent, exciting, familiar, favorable, flexible, flustered, friendly, frustrating, fun, good, hate, helpfulness, immediate, important, improving, inefficient, intelligent, interested, intuitive, involved, irritation, learnable, likable, lively, loved, motivating, natural, nice, personal, plain, pleasant, preference, presence, productive, quality, quick, relevant, reliable, respect, responsive, satisfied, sensate, sense of being together, sense of control, sense of success, simple, smooth, sociable, social presence, stimulating, successful, sufficient, surprising, time consuming, timely, tiring, trust, uncomfortable, understand, useful, user friendly, vexed, vivid, warm, and well-organized.

Such diversity may be seen as an expression of the cleverness of HCI researchers in assessing the different

aspects of subjective satisfaction of relevance to their work. Yet, comparisons across studies become difficult because of this diversity, as does the understanding of distinct dimensions of subjective satisfaction.

Another indication of the disarray is in the limited use of standardized questionnaires and the few studies that use measures that directly build upon earlier work. Of the 112 studies that measure satisfaction, 29 (26%) refer to previous work as a source for the questions. Among those, only twelve studies employ standardized questionnaires for measuring some kind of satisfaction. While specific studies assessing specific aspects of usability may justify the need for custom-made questions, a large group of studies do no more than measure what has above been called ease-of-use. Those studies add in their reinvention of ways to measure satisfaction little but lack of clarity and difficulties in comparing to other studies. Fifteen out of the 37 studies in the sample that measure ease-of-use do not even give the wording of all questions used, making this problem even more severe.

Finally, studies seldom refer to previous research in which particular questions have been used, and validation of the questions is seldom undertaken. Of the 112 studies that measure satisfaction, only 10 studies report validation efforts or measures of reliability, such as Cronbach's alpha.

A second comment on the satisfaction measures used is that studies vary greatly in the phenomena that are chosen for objective performance measures and those that are investigated by asking subjects about their perceptions and attitudes. One question arises when users' perception of phenomena is measured when those phenomena perhaps more fittingly could have been assessed by objective measures. McGrenere et al. (2002) is one example of such a study, investigating the customization of Microsoft Word. In that study, learnability was measured by asking subjects to indicate on a five-point Likert scale "overall satisfaction, ease of navigating through the menus and toolbars, control over MS Word, and ability to learn all the available features", (McGrenere et al., 2002, p. 167). But can learnability be measured in a valid way by asking users? Moreover, the abstract of the paper states,

> The study tested the effects of different interface designs on users' satisfaction and their perceived ability to navigate, control, and learn the interface. [...] Results showed that participants were better able to navigate through the menus and toolbars and were better able to learn with our prototype (McGrenere et al., 2002, p. 163).

These statements make a peculiar shift from mentioning perceived abilities to concluding by saying that learning was improved, and, as noted above, the validity of perceived ability to learn as an indicator of usability seems doubtful. In the context of McGrenere et al.'s study, this choice of measure may have been the only practical option or may have been what was considered most relevant for their particular research. Nevertheless, their study raises an important point more generally of how to make a sound choice among objective and subjective measures; we return to this issue in Section 4.

Third, the reader may have noted that the discussion above points to a class of measures that lie on the border between efficiency and satisfaction measures; that class is measures of perceived human effort. In the discussion of efficiency measures some questionnaires that assess subjective experiences of task difficulty were already mentioned (e.g., NASA's TLX). The ISO standard group measures of mental effort under efficiency (e.g. ISO, 1998, p. 13), while satisfaction measures include also "observation of overloading or underloading of the user's cognitive or physical workload" (ISO, 1998, p. 5), making an exclusive classification of measures difficult.

Fourth, the review shows that in practice subjective satisfaction is taken to mean a questionnaire completed after users used the interface. Only eight studies (4%) measure satisfaction during use without using questionnaires. Five of these look at the functions used by subjects to determine preference. In studying organization of photos, Rodden et al. (2001) forced users to chose which interface to use for finding photos, allowing users to switch between interfaces when they wanted. They then discuss the time users spent in each interface as an indicator of which interfaces users favored.

4. Challenges in measuring usability

In the previous section we reviewed the current practice of measuring usability and described some limitations with the measures employed. The aim of this section is, departing from the review and the problems discussed, to discuss challenges for how to conduct usability studies and for research into how to measure usability. In the final subsection, we attempt to communicate the challenges discussed in a model.

4.1. Subjective and objective measures of usability

In the studies reviewed some measures of usability concern users' perception of or attitudes towards the interface, the interaction, or the outcome. Let us call such measures *subjective* usability measures. Other measures concern aspects of the interaction not dependent on users' perception; on the contrary these measures can be obtained, discussed, and validated in ways not possible with subjective measures. Let us call these measures *objective* usability measures. In the literature on usability and performance measures, this distinction, while hard to define precisely, has been extensively used (Meister, 1985; Yeh and Wickens, 1988; Muckler and Seven, 1992). Note that by differentiating between objective and subjective measures, we do not attempt to make a substantial epistemological distinction. Such a distinction has been argued to simplify the nature of measurement in science (Muckler and Seven, 1992; Annett, 2002). Rather, we

suggest using the distinction to reason about how to choose usability measures and find more complete ways of assessing usability. The distinction may be applied to measures within all three aspects of the ISO (1998) standard discussed in previous sections.

One reason why we need study both objective and subjective measures of usability is that they may lead to different conclusions regarding the usability of an interface. Consider for example the case of objective time and subjectively experienced duration mentioned earlier. The study by Tractinsky and Meyer (2001) found a difference in subjectively experienced duration between interfaces when objective time was fixed. Similar differences have been exploited in work on subjective duration assessment (Czerwinski et al., 2001), which proposes as a new usability measure the ratio between objective time and subjectively experienced duration. Outside of HCI, psychologists have long recognized and quantified the difference between subjectively experienced duration and objective time (Eisler, 1976). Consequently, using both subjective and objective measures of time may give a more complete picture of usability, as differences between interfaces in objective time may not be found for subjectively experienced duration, and vice versa.

More general arguments for differences between objective and subjective measures may be found—besides observations from everyday life—in a variety of areas. Yeh and Wickens (1988), for example, outlined one theory for when subjective and objective measures of workload may dissociate. Another example is the meta-analysis of employer performance by Bommer et al. (1995) which found that objective and subjective ratings of employee performance had a mean correlation of .389, suggesting that these measures capture somewhat different aspects of performance. Such findings may hold also for some aspects of performance with computers.

As evident from the review, some studies mix together the very different measures of perceived learnability and changes in task efficiency, of subjective and objective assessment of outcomes, and of subjective and objective indicators of satisfaction. The distinction discussed above may serve to clarify how aspects of usability are described and reasoned about.

Another reason for pursuing the subjective/objective distinction is that for some aspects of usability we are interested not only in improving objective performance, but also in generating design advice on how to improve users' experience of interaction. Outside the HCI field, the relation between objective time and subjectively experienced duration has been discussed in relation to, for example, shop design (Underhill, 2000), where designs that lower buyers' experience of time passing (e.g., when waiting in line) have been experimented with. Among the reviewed studies, Tractinsky and Meyer (2001) make recommendations to designers of menu structures on how to lower users' experience of time based on the comparison between subjectively experienced duration and objective time. Conversely, in many contexts of use we are interested in limiting the objective time required to use an interface; indeed, a large number of the evaluations in the literature are undertaken with this interest. The utility of the subjective/objective distinction thus very much depend on the intended context of use. However, this distinction may help researchers and persons planning usability studies consider whether non-typical measures are relevant, in particular subjective measures of effectiveness (such as feelings of making high-quality work with office software), subjective measures of efficiency (such as feelings of quickly finding stuff to buy on e-commerce sites), and objective measures of satisfaction (such as physiological measures of fun in playing computer games). Depending on the context, a balanced focus on subjective and objective measures may help improve both the user experience and objective performance.

In summary the challenges to research we see are to develop subjective measures for aspects of quality-in-use that are currently mainly measured by objective measures, and vice versa, and evaluate their relation. Such developments seem especially to be lacking for outcome quality vs. perceived outcome, time vs. subjective experienced duration, perceived learnability vs. changes in time to complete tasks, and objective measures of satisfaction vs. subjective satisfaction questionnaires. In studies of usability, we suggest paying special attention to whether subjective or objective measures are appropriate, and whether a mix of those two better covers the various aspects of quality-in-use. Practitioners should take care when reasoning about usability not to conflate subjective and objective measures.

4.2. Measures of learnability and retention

To reflect on the usability measures classified in Section 3, in particular measures of efficiency, we find it relevant to compare them to recommendations on how to measure usability. The intent of this comparison is to start a discussion of the completeness of the measures employed. The well-known textbook by Ben Shneiderman (1998, p. 15) recommends measuring (1) time to learn, (2) speed of performance, (3) rate of errors by users, (4) retention over time, and (5) subjective satisfaction. Nielsen (1993, p. 26) similarly recommends measuring (a) learnability, (b) efficiency, (c) memorability, (d) errors, and (e) satisfaction.

Most of the reviewed studies follow part of the recommendations by measuring task completion time (points 2 and b above), accuracy (points 3 and d), and satisfaction with the interface (points 5 and e): 92% of the studies measure at least one of these; 13% of the studies measure all three. However, as mentioned earlier, learnability (points 1 and a), for example the time it takes to learn an interface, is only measured in five studies; retention (points 4 and c) appears only to be directly measured and discussed in one study. In Isokoski and Käki (2002) the ability to learn input devices was measured as the decrease in task completion time as users become

experienced with the devices. Each user completed 20 sessions each lasting approximately 35 min. This allowed Isokosi and Käki to draw curves of how the time spent on entering one digit decreases over time, facilitating a discussion of differences between the learnability of interfaces. In the study by Czerwinski et al. (1999), users' ability to come back to and effectively use an organization of web pages after a delay of 4 months was tested, addressing the aspect of retention mentioned above. However, the majority of studies make no attempt to measure learnability or retention.

Note that we are talking about measures of learning based on users' interaction with the interface. Another approach is used by Wulf and Golombek (2000), who asked users to answer questions where they must explain or predict the functioning of a tailorable groupware application after having used the application. Examples of these questions are "The user 'golombek' sends you a document template. How can you perceive this fact?" and gives the user a list of five possible answers including "an icon is displayed in the status bar because …" and "the document template will be in my mail box" (p. 263). The relation between questionnaire items on learnability and, for example, changes in task completion times over prolonged use, has not as far as we know been studied enough to warrant employing only questionnaire-based (or subjective) measures of learnability.

This challenge is most relevant for studies or research addressing systems that users should be able to learn quickly or that will be intensively used (and where a steep learning curve may be acceptable if users eventually learn to become skilled with the system). Also, for systems aimed at intermittent use it seems relevant to consider the challenges related to retention mentioned above. For walk-up-and-use systems, this challenge may not matter at all because the learnability of such systems may be uncovered in standard usability tests.

Overall, usability studies could put more emphasis on measures of learning, for example by measuring the time needed to reach a certain level of proficiency. In addition, measures of the retention of objects and actions available in the interface (i.e., the ability of users to come back and successfully use the interface) are important in gaining a more complete picture of usability. Without these, and with the common use of one-session studies only, we know little about usability of interfaces that are used repeatedly. Research should focus on developing easy-to-adopt techniques for measuring learning and retention.

4.3. Measures of usability over time

The studies reviewed show that users typically interact only briefly with interfaces under investigation; as mentioned earlier the median duration of users' interaction was 30 min; only 13 studies examined interaction that lasts longer than five hours. As a representative example of a long-term study, consider that of McGrenere et al. (2002). Over approximately six weeks, they studied how participants used a customizable version of Microsoft Word by logging interactions and by administering seven questionnaires. This allows for some reasoning about how experienced users utilize customization.

The brief period of interaction in the studies reviewed explains the lack of focus on measures of learning and retention discussed above. Not only does this observation show that longer-term studies are rare, it also suggests that we have little quantitative evidence about what long-term usable systems are like. The observation also suggests that we know little about how usability develops as the user spend more time interacting with the interface and how tradeoffs and relations between usability aspects change over time. In particular, it would be relevant to know more about how measures of effectiveness and satisfaction develop over time.

This challenge appears to hold broadly, except for walk-up-and -use systems and for systems when the expected duration of users' interaction is similar to or less than the median time of common usability studies. For use contexts where longer interaction is expected, we need to consider whether the results of our snapshot usability studies remain relatively constant over time.

In summary, when conducting usability studies we could consider studying how usability developers over time, especially over periods of time exceeding what is commonly studied in HCI. This could naturally include changes in efficiency and effectiveness over time as one indication of how users learn to use the interface. From research, we need a more full understanding of how the relation between usability aspects develops over time. Do usability measures converge over time in pointing out a particular interface as superior to other interfaces? Are users able, over time, to compensate for most usability problems that lead to initial dissatisfaction?

4.4. Extending, validating and standardizing measures of satisfaction

The disarray of measures of satisfaction presents special challenges. One is to extend the existing practice of measuring satisfaction almost exclusively by post-use Likert-scale questions; another is to validate and standardize the questions used.

Almost all satisfaction measures are obtained by questionnaires administered to users after they have used a system—93% of the studies used exclusively measures of this kind. Among the problems with questionnaires are that they are collected after the fact, that they are shaped by individuals' (mis)understanding of the questions, and that they provide general information that is hard to link to specific parts of the interaction or the interface. The studies reviewed include notable attempts at extending satisfaction measures to avoid these problems. Nichols et al. (2000) investigated the impact of different virtual environments (VE) on the experience of presence. Presence was assessed

with a set of seven-point rating scales that subjects completed after using the VEs. In addition, and of particular interest here, Nichols et al. measured presence also as a reflex response: "The VE was programmed with a randomly timed 'startle event' and the participants' reactions were classified into three categories—no reaction, a verbal report of a reaction, or a physically noticeable reaction" (p. 476). The participants' reactions to startle events allow for non-questionnaire assessment of one aspect of presence. Another example is a study by Izsó and Láng (2000) that investigated heart period variability as an indicator of mental effort. Izsó and Láng showed that heart period variability could predict users' answers to post-task questions about task complexity.

Outside the studies reviewed similar efforts have been undertaken. Tattersall and Foord (1996) presented a technique for collecting users' satisfaction ratings during rather than after use of a system, and experiments with physiological measures of usability have also been undertaken (Wastall, 1990; Mullins and Treu, 1991; Allanson and Wilson, 2002). The above examples appear as one interesting research direction for exploring supplements to post-use satisfaction questionnaires.

The second challenge with respect to satisfaction measures is to validate and, where possible, work out standards for such measures. For many constructs measured in the studies reviewed, validated questionnaires are available: constructs such as anxiety (Bailey et al., 2001), presence (Sällnas et al., 2001), and self-disclosure (Dahlbäck et al., 2001) were assessed with standardized questionnaires. Other studies achieve some degree of validity by building upon measurement instruments discussed in previous work, such as for entertainment (O'Keefe et al., 2000) or for subjective responses to computer-mediated conversation (Garau et al., 2001). However, measures of ease-of-use are reinvented again and again, despite the availability of several validated questionnaires such as Chin et al. (1988) and Davis (1989). The measures of specific attitudes towards the interface used also seem partly covered by existing questionnaires. SUMI (Kirakowski and Corbett, 1993), for example, includes subscales of control and learnability; QUIS (Chin et al., 1988) contains a series of questions assessing how easily users can learn to use the system. Table 5 shows that several studies have been trying to measure these aspects of usability. The utility of studies that reinvent measures or explore satisfaction measures without any assessment of reliability or validity is doubtful.

Some readers may be skeptical about the use of standardized questionnaires because they feel that such questionnaires are not applicable in the context-of-use under consideration or are unnecessarily limiting the scope of studies of usability. While we acknowledge that questionnaires for all constructs of interest to HCI do not exist, skeptical readers may appreciate that comparing studies using standardized questionnaires would be easier than comparing the studies reviewed. Also, the results from studies using such questionnaires may be taken up with greater confidence.

This challenge seems to be relevant for almost all the use contexts where we routinely would measure some aspect of subjective satisfaction. In particular, when the specific aspects of satisfaction that are relevant do not include overall satisfaction—but for example fun, trust, or flow—it seems especially relevant to consider if we can validate or use a standard for the particular measure.

In summary, persons who conduct usability studies are well advised to use standardized questionnaires whenever possible. Such questionnaires are available both for overall satisfaction and for specific attitudes. A challenge to research on usability is to extend satisfaction measures beyond post-use questionnaires, and to focus more on validation and standardization of satisfaction measures. Validation may be achieved through studies of correlation between measures to be discussed next.

4.5. Studies of correlations between measures

A weak understanding of the relation between usability measures gives rise to many of the issues discussed in this review. With a better understanding, we could make more informed choices about which usability measures to employ. Studies of correlation between measures may improve this understanding by informing us whether our measures contribute something new and what their relation are to other aspects of usability. Such studies appear to be needed within all of the ISO categories of usability aspects, but also between aspects, so as to characterize what measures of a particular aspect (e.g., efficiency) contribute, that are not captured by measures of other aspects (e.g., effectiveness).

The studies reviewed contain several interesting observations based on studies of correlations between usability measures. Karat et al. (2001) studied the correlation between mouse activity and satisfaction measures in the context of a web application intended to give entertaining access to multimedia. Correlations suggested that less clicking leads to more watching and consequently more engaging and entertaining web experiences. In a study of accuracy measures for pointing devices, MacKenzie et al. (2001) used correlations to investigate how seven proposals for new usability measures were related to throughput. From the correlations obtained, they identified promising measures.

Outside the studies reviewed, papers discuss correlations between different aspects of usability (Frøkjær et al., 2000; Hassenzahl et al., 2000). The main contribution of these papers seems to be their challenge of what should be measured in usability studies. What seems a mostly unexplored benefit of studies of correlations between measures is getting a better understanding of not only *if* measures are correlated or not, but also *when* or under what conditions measures are correlated. For example, given that satisfaction is not always correlated with

effectiveness (as argued by Frøkjær et al., 2000), what does this signify in a particular context? Are there critical aspects of effectiveness we are ignoring? Of satisfaction? Are we looking at too short-term use? These questions seem worth exploring.

This challenge appears mainly relevant for usability research. There we see the need for a better understanding of the relation between usability measures, for which studies of correlations between measures would be one contribution. Such studies could shed light on the earlier discussions that usage patterns seem an insecure indicator of quality-in-use, and help investigate when objective and subjective usability uncover different aspects of usability. For new measures of usability suggested to be of crucial importance in emerging contexts of use, correlations between measures seem especially important to justify the relevance and necessity of new measures. In the long term, persons who conduct usability studies could use such studies to inform their choice of usability measures; at present, they should strongly consider using individual measures of aspects of usability that are often unrelated (e.g., satisfaction and effectiveness).

4.6. Micro and macro measures of usability

As may have struck the reader, the same measure of usability may be classified differently according to what level a task is considered at. For example, in studies of input devices, users' accuracy in rotating of 3d objects is often used as an effectiveness measure (e.g., Partala, 1999). However, rotation forms part of some higher-level tasks that are usually more cognitively or socially complex and have a longer duration. For example, rotation may be part of tasks such as learning about 3d shapes, the support for which could be assessed by measures of recall (e.g., recall of shapes interacted with) or as the effectiveness in completing tasks for which understanding of shapes are needed (e.g., drawing of 3d models).

So, the understanding of usability measures depend on the level at which tasks are considered. A crude way to think of this is to consider two levels of tasks on a continuum. One concerns tasks and measures of usability at a *micro level*. Such measures cover tasks that are usually of short duration (seconds to minutes), has a manageable complexity (most people will get them right), often focus on perceptual or motor aspects (visual scanning, mouse input), and time is usually a critical resource. Another concerns tasks and measures of usability at a *macro level*. Such measures cover tasks that are longer (hours, days, months), are cognitively or socially complex (require problem-solving, learning, critical thinking, or collaboration), display large individual differences in the interaction process and vast variations in the outcome, and usually have effectiveness and satisfaction as critical parameters.

Fig. 1 illustrates this distinction. Given that tasks have different durations, as indicated by the left-most scale, they can be seen as having mainly either a macro or micro level focus (second column in the figure). Within either focus, certain characteristics of the task will be in focus, for example whether its solutions will display variability or uniformity, or whether it concerns perceptual or social issues (third column in the figure). Given these foci, certain measures of evaluation follow naturally, for example to focus on task completion times or to investigate quality of the work products created (right-most column in the figure). Note that time does not in itself determines the micro/macro distinction, but is closely related to it. However, the micro-macro distinction appears to be most useful for choosing measures of satisfaction and of effectiveness.

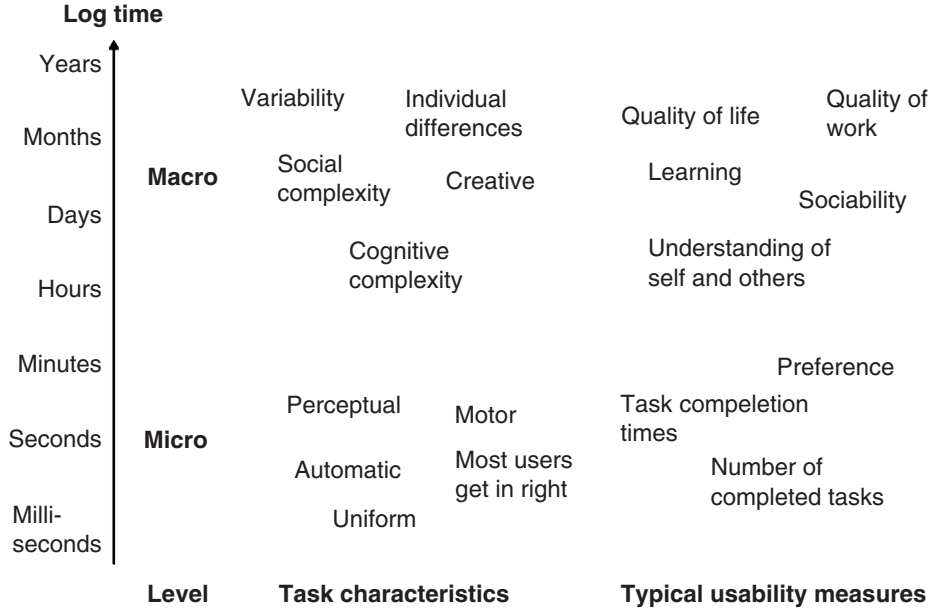

Fig. 1. Micro and macro measures of usability.

In the studies reviewed, this simple distinction leads to a few important observations. First, the macro perspective on tasks is rare; we seem most often to cope with the complexity of usability evaluation by choosing simple, manageable measures at a micro level. As mentioned above, only a handful of the reviewed studies investigate long-term interaction. Additionally, few studies use tasks that allow quality of outcome and learning to be measured (see Table 2). Yet, the focus on micro tasks and the belief that we can safely decompose most macro tasks into micro tasks, and reason about usability based on micro tasks only, seems dubious. The grand goals of user interfaces that stimulate creativity (Shneiderman, 2000), support sociability on the internet (Preece, 2000), enables personal fulfillment, and so forth, seem unlikely to be evaluated, let alone achieved, if we focus on micro measures, as these goals seem to involve psychological and social complexities only visible in macro tasks.

Second, the earlier discussed differences in interpreting usability measures seem to stem from the fact that the same measures have different interpretations according to whether one considers micro or macro tasks. Short task completion times, for example, are crucial in studies of input devices; for creativity–support interfaces or educational software, high task completion times may be seen as indicative of motivation and engagement, as has been the case in some of the reviewed studies (e.g. by Inkpen, 2001, see Section 3.2).

The challenge to research we see here is to explore more thoroughly the relation between micro and macro measures of usability. This challenge appears to be valid across a variety of system types and use contexts, except when we—by nature of the system, persons or tasks studied—are only interested in macro-level tasks (e.g., tools for relationship building) or micro-level tasks (e.g., many studies of input techniques). For usability studies, we suggest to focus, where relevant and practically feasible, on macro tasks. In many cases, this will force us towards more ambitious goals and more challenging usability studies.

4.7. A working model for usability measures and research challenges

In an attempt to summarize the challenges discussed above and to highlight some usability measures we find especially important, we propose Fig. 2 as a working model of usability measures and research challenges. The reader should note three things about the model.

First, this model suggests questions and measures to consider when selecting usability measures for a study. On the figure, six categories of measures are shown in bold. These categories have been found in the review to be of particular importance (e.g., validated questionnaires) or requiring particular care in their interpretation (e.g., usage patterns). In conducting usability studies, it are useful to check whether these measures may be relevant in the particular context investigated, whether measures from all six categories can be obtained and are useful, and whether the questions in italic inspire selecting more valid or complete usability measures. If the measures listed in the figure are for some reason irrelevant, Tables 2–5 contain further inspiration for finding measures.

Second, the model show in italic research questions related to usability measures. These questions have been put forward and justified in previous subsections. Lack of research addressing them appears to be one reason for the somewhat dissatisfying state of affairs identified in Section 3.

Third, the model can be seen as comprising a number of improvements over the ISO, 9241-part 11 standard for usability (ISO, 1998). We have changed some terminology that in the preceding discussions have been found unclear or improvable. For effectiveness measures we find it useful to talk instead of measures of the quality or satisfaction

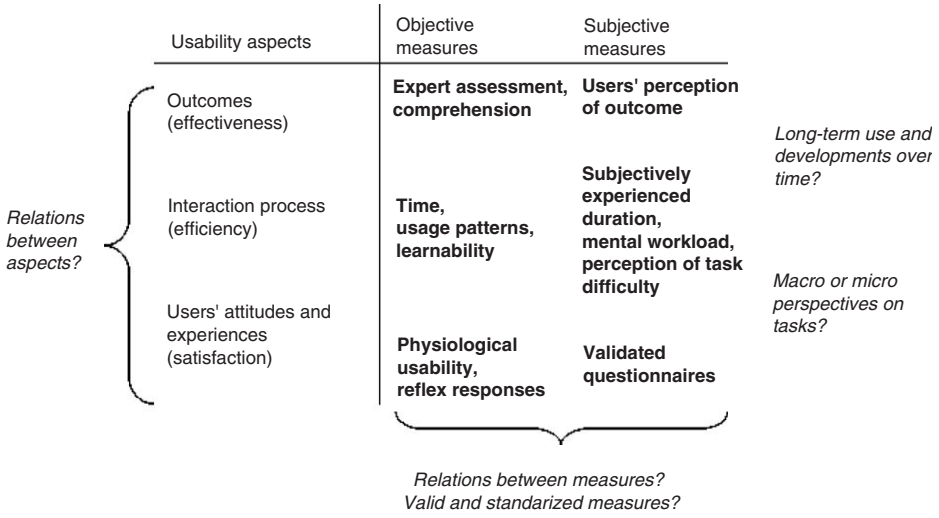

Fig. 2. Challenges in measuring usability.

with *outcomes* of the interaction. In relation to the ISO model, we also include users' perception of work products and perceptions of whether they reached the intended outcomes or not. As discussed in Section 3, with the ISO standard one can equally well consider such measures as indicators of satisfaction. Note that outcomes includes both tangible outcomes, such as work products, and intangible ones, for example changes in attitudes, having fun, or improving relations to other persons. Focusing on outcomes also means that some measures suggested in the ISO standard have to be rethought. For example, the suggestion that "[p]ercentage of relevant functions used" (p. 11) is an effectiveness measures is clearly inaccurate; rather, this appear to be measure of the interaction process. Likewise with "[n]umber of references to documentation" (p. 11), which also seems to concern the interaction process.

The ISO standard also suggests that effectiveness measures may be combined from measures of accuracy and completeness (ISO, 1998, p. 12). Yet, this repeats the problems discussed above on combining measures and we recommend reporting such measures separately.

For efficiency measures we find the ISO definition ("resources expended in relation to the accuracy and completeness with which users achieve goals", p. 2) to mix together effectiveness and efficiency measures. As discussed in Section 3, we suggest not to involve accuracy and completeness in the calculation of efficiency, but always to report efficiency measures per task or goal. In addition to making the results of usability studies easier to interpret, this suggestion also keeps measures independent. Instead of using the term efficiency, as in ISO, it thus makes more sense to talk about measures of the *interaction process*, understanding that the focus here is on the process of users' interacting with the interface to achieve the intended outcome. In any case, we differentiate between subjectively experienced and objectively measured aspects of the interaction process.

For satisfaction measures we find it more useful to talk about measures of *users' attitudes and experience*, rather than restricting satisfaction to "freedom from discomfort, and positive attitudes towards the use of the product" (ISO, 1998, p. 2). The focus here is on users attitudes towards the interface and their experience of using the interface. Thus, these measures are not about the outcome, nor about the process of interaction.

In relation to the question of how to choose usability measures, we believe that Fig. 2 introduces important distinctions compared to the ISO standard. In ISO, for example, it is claimed that "[i]f it is not possible to obtain objective measures of effectiveness and efficiency, subjective measures based on the user's perception can provide an indication of effectiveness and efficiency" (p. 5). Section 3 illustrates how such mixing together of subjective measures may lead to unclear, and perhaps invalid, conclusions. In the ISO standard (e.g., p. 10) measures such as "[r]ating scale for satisfaction" and "[f]requency of discretionary use" are also mixed together. On the contrary, we suggest keeping distinct objective and subjective measures of various usability aspects. This is also suggested in an appendix to the standard (ISO, 1998, p.12) but not reflected in many of the discussions within the standard itself.

The ISO standard gives some guidance in selecting tasks ("[f]or the purposes of evaluating usability, a set of key tasks will typically be selected to represent the significant aspects of the overall task", p. 4). However, the insistence on using macro tasks we believe will lead towards bolder and more challenging measures.

While this model provides a number of benefits, for example over the ISO model, we believe that the separation of outcome, process, and attitude measures may need even further explanation and criteria for distinguishing measures. It should also be noted that a substantial group of measures concern users' attitudes towards other phenomena than the interface, notably other persons. While we did not let such factors into the model as a distinct usability aspect, measures of relations to other people, for example, are often relevant and could add a new, fourth aspect of usability.

5. Conclusion

We have reviewed usability measures employed in 180 studies published in core HCI journals and proceedings, summarizing current practice of measuring usability as well as critically reviewing that practice. Notable problems in how usability measures are employed include (1) measures of the quality of interaction, for example assessed by domain experts, are used only in a few studies; (2) approximately one quarter of the studies do not assess the outcome of the users' interaction, leaving unsupported any broad claims about usability; (3) measures of learning and retention of how to use an interface are rarely employed, despite being recommended in prominent textbooks; (4) some studies treat measures of how users interact with interfaces as being synonymous with quality-in-use despite an unclear, if not weak, relation between usage patterns and quality-in-use; (5) measures of users' satisfaction with interfaces are in a disarray and most studies reinvented questions to be asked users, ignoring validated questionnaires readily available; and (6) some studies mix together, perhaps even consider synonymous, users' perceptions of phenomena (e.g., the learnability of an interface) with objective measures of those phenomena (e.g., time needed to master an interface to a certain criterion).

Based on the review, we proposed several challenges with respect to measuring usability. Those challenges include the need to understand better the relation between objective and subjective measures of usability; to understand better how to measure learnability and retention; to extend satisfaction measures beyond post-use questionnaires; to study correlations between measures; and to push the

boundaries of what we conceive as usability measures by focusing on macro measures, such as those related to cognitively and socially complex tasks, and long-term use. In addition, we proposed a working model claimed to embody several improvements over the ISO standard for usability measures.

The reader should keep in mind two limitations of the paper. First, we only analysed research studies and did not address problems and challenges of usability testing in the software industry. It is unclear if the discussions in this paper generalize to this setting; a separate review of industry usability practice is needed to investigate this. Another limitation is that our discussions have not attempted in detail to account for the measures used in different contexts or how different tasks and domains impact the choice of usability measures. We do not suggest that usability can be fully accounted for outside of a particular context of use; the arguments of Newman and Taylor (1999) convincingly suggest otherwise. However, our review indicates that even discussing and analysing usability measures at a general level can identify problems and research challenges concerning how to measure usability in particular contexts of use.

Despite more than 20 years of research into usability, current practice in measuring usability suggests that choosing usability measures is difficult and that the conclusions of some usability studies are weakened by their choices of usability measures and by the way they use measures of usability to reason about quality-in-use. The suggestions for how to meet the challenges identified may, if pursued in research and implemented in usability studies, make choosing usability measures easier and establish more valid and complete usability measures.

Acknowledgments

From the spring of 1999 and on, I have benefited greatly from discussions with Erik Frøkjær and Morten Hertzum of issues related to this paper. Erik Frøkjær and Aran Lunzer provided helpful comments on an early draft.

Appendix A. Studies included in the review

A.1. Papers in ACM Transactions on Human–Computer Interaction

Ren, X. and Moriya, S. (2000), 7 (3), 384–416; Basdogan, C., et al. (2000), 7 (4), 443–460; Sallnäss, E.-L. et al. (2000), 7 (4), 461–476; Inkpen, K. (2001), 8 (1), 1–33; Sedig, K. et al. (2001), 8 (1), 34–59; Suhm, B. et al., 8 (1), 60–98; Hornof, A. (2001), 8 (3), 171–197; Thomas, B. and Calder, P. (2001), 8 (3), 198–222.

A.2. Papers in Behavior & Information Technology

Tractinsky, N. (2000), 19 (1), 1–13; Alsio, G. and Goldstein, M. (2000), 19 (2), 87–96; Mead, S. et al. (2000), 19 (2), 107–123; Carayon, P. and Karsh, B.-T. (2000), 19 (4), 247–262; Izsó, L. and Láng, E. (2000), 19 (4), 297–306; Sauer, J. et al. (2000), 19 (5), 315–327; Lim, J. (2000), 19 (5), 329–338; Matarazzo, G. and Sellen, A. (2000), 19 (5), 339–348; Healey, C. (2000), 19 (5), 349–366; Schenkman, B. and Jönsson (2000), 19 (5), 367–377; Tanin, E. et al. (2000), 19 (6), 393–403; Jacko, J. et al. (2000), 19 (6), 427–439; Zimmermann, C. and Bridger, R. (2000), 19 (6), 441–449; Norman, K. et al. (2001), 20 (1), 37–45; Labiale, G. (2001), 20 (3), 149–158; Sears, A. et al. (2001), 20 (3), 159–166; Marshall, D. et al. (2001), 20 (3), 167–188; Castelhano, M. and Muter, P. (2001), 20 (4), 237–247; Wulf, V. and Golombek, B. (2001), 20 (4), 249–263; Xie, X. et al. (2001), 20 (4), 281–291; MacKenzie, S. and Zhang, S. (2001), 20 (6), 411–418; Westerman, S. et al. (2001), 20 (6), 419–426; Lind, M. et al. (2001), 20 (6), 427–432.

A.3. Paper in Human–Computer Interaction

Trafton, G. and Trickett, S. (2001), 16 (1), 1–38.

A.4. Papers in International Journal of Human–Computer Studies

Burns, C. (2000), 52 (1), 111–129; Anderson, A., et al., 52 (1), 165–187; Toms, E. (2000), 52 (3), 423–452; Nichols, S. et al. (2000), 52 (3), 471–491; Kim, J. and Yoo, B. (2000), 52 (3), 531–551; Tyfa, D. and Howes, M. (2000), 52 (4), 637–667; Whittle, J. and Cumming, A. (2000), 52 (5), 847–878; Fang X. and Salvendy, G: (2000), 52 (5), 915–931; Monk, A. and Watts, L. (2000), 52 (5), 933–958; Cutmore, T. et al. (2000), 53 (2), 223–249; Isbister, K. and Nass, C: (2000), 53 (2), 251–267; Head, M. (2000), 53 (2), 301–330; Ruddle, R. et al. (2000), 53 (4), 551–581; O'Keefe, R. et al. (2000), 53 (4), 611–628; Risden, K. et al. (2000), 53 (5), 695–714; Westerman, S. and Cribbin, T. (2000), 53 (5), 765–787; Stasko, J. et al. (2000), 53 (5), 663–694; North, C. and Shneiderman, B. (2000), 53 (5), 715–739; Batra and Antony, S. (2001), 54 (1), 25–51; Kontogiannis, T. and Linou, N. (2001), 54 (1), 53–79; Gregor, S. (2001), 54 (1), 81–105; Kehoe, C. et al. (2001), 54 (2), 265–284; Tung, S.-H. et al. (2001), 54 (3), 285–300; Tang, H. (2001), 54 (4), 495–507; Kettanurak, V. et al. (2001), 54 (4), 541–583; Dyson, M. and Haselgrove, M. (2001), 54 (4), 585–612; Hone, K. and Baber, C. (2001), 54 (4), 637–662; France, E. et al. (2001), 54 (6), 857–876; Goonetilleke, R. et al. (2001), 55 (5), 741–760; Tractinsky, N. and Meyer, J. (2001), 55 (5), 845–860.

A.5. Papers in ACM CHI Conference

Accot, J. and Zhai, S. (2001), 1–8; MacKenzie, S. et al. (2001), 9–16; Duh, H. et al. (2001), 85–89; Gray, W. and Fu, W.-T. (2001), 112–119; Mamykina, L. et al. (2001), 144–151; Gong, L. and Lai, J. (2001), 158–165; Baldis, J. (2001), 166–173; Rodden, K. et al. (2001), 190–197; Woodruff, A. et al. (2001), 198–205; Lai, J. et al. (2001),

206–212; Buyukkokten, O. et al. (2001), 213–220; Corbett, A. and Anderson, J. (2001), 245–252; Dumais, S. et al. (2001), 277–284; Hornbæk, K. and Frøkjær, E. (2001), 293–300; Garau, M. et al (2001), 309–316; Wang, J. et al. 349–356; James, C. and Reischel, K. (2001), 365–371; Bickmore, T. and Cassell, J. (2001), 396–403; Tan, D. et al. (2001), 418–425; Mason, A. et al. (2001), 426–433; Cockburn, A. and McKenzie, B. (2001), 434–441; Liu, Q. et al. (2001), 442–449; Rui, Y. et al. (2001), 450–457; Shoemaker, G. and Inkpen, K. (2001), 522–529; Zhai, S. et al. (2002), 17–24; Isokoski, P. and Käki, M. (2002), 25–32; Myers, B. et al. (2002), 33–40; McGuffin, M. and Balakrishnan, R. (2002), 57–64; Hinckley, K. et al. (2002), 65–72; Accot, J. and Zhai, S. (2002), 73–80; Pierce, J. and Pausch, R. (2002), 105–112; Jettmar, E. and Nass, C. (2002), 129–134; Bos, N. et al. (2002), 135–140; Zheng, J. et al. (2002), 141–146; Scott, S. et al. (2002), 155–162; McGrenere, J. et al. (2002), 164–170; Czerwinski, M. et al. (2002), 195–202; Cockburn, A. and McKenzie, B. (2002), 203–210; Ehret, B. (2002), 211–218; Drucker, S. et al. (2002), 219–226; Baudisch, P. et al. (2002), 259–266; Gutwin, C. (2002), 267–274; Whittaker, S. et al. (2002), 275–282; Suhm, B. et al. (2002), 283–290; Pirhonen, A. et al. (2002), 291–298; Terveen, L. et al. (2002), 315–322; Jacob, R. et al. (2002), 339–346; Bayles, M. (2002), 363–366; Farnham, S. et al. (2002), 375–382; McGee, D. et al. (2002), 407–414; Robertson, G. et al. (2002), 423–430.

A.6. Papers in IFIP Interact Conference

Stolze, M. (1999), 45–53; Senda, Y. et al. (1999), 102–109; Baber, C. et al. (1999), 126–133; Halverson, C. et al. (1999), 133–140; Golightly, D. et al. (1999), 149–155; Czerwinski, M. et al. (1999), 163–170; Hornbæk, K. and Frøkær, E. (1999), 179–186; Girgensohn, A. et al. (1999), 205–212; Geissler, J. et al. (1999), 222–230; Bérard, F. (1999), 238–244; Mitsopoulos, E. and Edwards, A. (1999), 263–271; Beveridge, M. and Crerar, A. (1999), 272–280; Leung, Y. and Morris, C (1999), 287–294; McFarlane, D. (1999), 295–303; Anderson, A. et al. (1999), 313–320; Dulberg, M. (1999), 375–382; Campbell, C. et al. (1999), 383–390; Wang, Y. and MacKenzie, C. (1999), 391–398; Girgensohn, A. et al. (1999) 458–465; Partala, T. (1999), 536–543; Norris, B. (1999), 544–551; Cockayne, A. et al. (1999), 582–588; Vetere, F. and Howard, S. (1999), 589–596; Barker, T. et al. (1999), 648–555; Smith, B. and Zhai, S. (2001), 92–99; Partala, T. et al. (2001), 100–107; Hashimoto, W. and Iwata, H. (2001), 108–114; Farnham, S. et al. (2001), 115–122; McCrickard, S. et al. (2001), 148–156; Bartram, L. et al. (2001), 157–165; Cribbin, T. and Chen, C. (2001), 166–173; Pacey, M. and MacGregor, C. (2001), 174–181; Takahashi, T. et al. (2001), 190–197; Fjeld, M. et al. (2001), 214–223; Lehikoinen, J. (2001) 224–231; Hourizi, R. and Johnson, P. (2001), 255–262; Cutrell, E. et al. (2001), 263–269; Suzuki, N. et al. (2001), 278–285; Dahlback, N. et al. (2001), 294–301; Poon, J. and Nunn, C. (2001), 302–309; Wang, Q. et al. (2001), 310–317; Brumitt, B. and Cadiz, JJ (2001), 375–382; Nyberg, M. et al. (2001), 391–398; Baber, C. et al. (2001), 439–447; Karat, C.-M. et al. (2001), 455–462; LeeTiernan, S. and Grudin, J. (2001), 472–479; Takeuchi, Y. et al. (2001), 480–487; Vanhoucke, V. et al. (2001), 530–536; Hertzum, M. et al. (2001), 537–544; Dieberger, A. and Russell, D. (2001), 545–552; Leung, Y. et al. (2001), 553–560; Bailey, B. et al. (2001), 593–601; Pollard, N. and Monk, A. (2001), 602–608.

References

Allanson, J., Wilson, G.M., 2002. Workshop on physiological computing. Extended abstracts of ACM Conference on Human Factors in Computer Systems. ACM Press, New York, NY, pp. 912–913.

Amabile, T.M., 1996. Creativity in context. Westview Press, Boulder, CO.

American Psychological Association, 1985. Standards of Psychological Testing. American Psychological Association, Washington, DC.

Annett, J., 2002. Subjective rating scales science or art? Ergonomics 45 (14), 966–987.

Bailey, B.P., Konstan, J.A., Carlis, J.V., 2001. The effects of interruptions on task performance, annoyance, and anxiety in the user interface. In: Proceedings of IFIP TC.13 International Conference on Human–Computer Interaction. IOS Press, Amsterdam, pp. 593–601 *.

Basdogan, C., Ho, C., Srinivasan, M.A., Slater, M., 2000. An experimental study of the role of touch in shared virtual environments. ACM Transactions on Human–Computer Interaction 7 (4), 443–460 *.

Bayles, M., 2002. Designing online banner advertisements: should we animate? In: Proceedings of CHI 2002 ACM Conference on Human Factors in Computer Systems. ACM Press, New York, NY, pp. 363–366 *.

Bennett, J.L., 1972. The user interface in interactive systems. Annual Review of Information Science 7, 159–196.

Bennett, J., 1979. The commercial impact of usability in interactive systems. 1–17. In: Shackel, B., Man/computer Communication: Infotech State of the Art Report, vol. 2. Infotech International, Maidenhead, UK.

Bevan, N., 1995. Measuring usability as quality of use. Software Quality Journal 4, 115–150.

Bommer, W.H., Johnson, J.L., Rich, G.A., Podsakoff, P.M., Mackenzie, S.B., 1995. On the interchangeability of objective and subjective measures of employee performance: a meta-analysis. Personnel Psychology 48, 587–605.

Bos, N.D., Judy, S.O., Darren, G., Gary, M.O., 2002. Effects of four computer-mediated channels on trust development. In: Proceedings of ACM Conference on Human Factors in Computer Systems. ACM Press, New York, NY, pp. 135–140 *.

Brickmore, T., Cassell, J., 2001. Relational agents: a model and implementation of building user trust. In: Proceedings of ACM Conference on Human Factors in Computer Systems. ACM Press, New York, NY, pp. 396–403 *.

Burns, C.M., 2000. Navigation strategies with ecological displays. International Journal of Human–Computer Studies 52 (1), 111–129.

Card, S., Moran, T., Newell, A., 1983. The Psychology of Human–Computer Interaction. Lawrence Erlbaum, Hillsdale, NJ.

Carroll, J., Thomas, J., 1988. Fun. SIGCHI Bulletin 19 (3), 21–24.

Chin, J.P., Diehl, V.A., Norman, K.L., 1988. Development of an instrument for measuring user satisfaction of the human–computer interface. In: Proceedings of ACM Conference on Human Factors in Computing Systems. ACM Press, New York, NY, pp. 213–218.

Cockburn, A., McKenzie, B., 2001. 3D or Not 3D? Evaluating the effect of the third dimension in a document management system. In:

*Denotes a reference among the reviewed studies.

Proceedings of ACM Conference on Human Factors in Computing Systems. ACM Press, New York, NY, pp. 434–441 *.

Cook, T.D., Campbell, D.T., 1979. Quasi-Experimentation: Design and Analysis Issues for Field Settings. Rand McNally College Publishing Company, Chicago, IL.

Corbett, A.L., Anderson, J., 2001. Locus of feedback control in computer-based tutoring: impact on learning rate, achievement and attitudes. In: Proceedings of ACM Conference on Human Factors in Computer Systems. ACM Press, New York, NY, pp. 245–252 *.

Cribbin, T., Chen, C., 2001. Exploring cognitive issues in visual information retrieval. In: Proceedings of IFIP TC.13 International Conference on Human–Computer Interaction. IOS Press, Amsterdam, pp. 166–173 *.

Czerwinski, M.P., van Dantzich, M., Robertson, G., Hoffman, H., 1999. The contribution of Thumbnail image, mouse-over text and spatial location memory to web page retrieval in 3D. In: Proceedings of IFIP TC.13 International Conference on Human–Computer Interaction. IOS Press, Amsterdam, pp. 163–170 *.

Czerwinski, M., Horvitz, E., Cutrell, E., 2001. Subjective duration assessment: an implicit probe for software usability? In: , Proceedings of IHM-HCI 2001 Conference, vol. 2. Cépaduès-Editions, Toulouse, France, pp. 167–170 *.

Dahlbäck, N., Swamy, S., Nass, C., Arvidsson, F., Skågeby, J., 2001. Spoken interaction with computer in native or non-native language—same of different? In: Proceedings of IFIP TC.13 International Conference on Human–Computer Interaction. IOS Press, Amsterdam, pp. 294–301 *.

Davis, F.D., 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 13 (3), 319–340.

Dillon, A., 2001. Beyond usability: process, outcome and affect in human computer interactions. Canadian Journal of Information Science 26 (4), 57–69.

Dix, A., Finlay, J., Abowd, G., Beale, R., 1993. Human–computer interaction. Prentice-Hall, New York, NY.

Drucker, S.M., Glatzer, A., De Mar, S., Wong, C., 2002. SmartSkip: consumer level browsing and skipping of digital video content. In: Proceedings of ACM Conference on Human Factors in Computer Systems. ACM Press, New York, NY, pp. 219–226 *.

Dumais, S.T., Cutrell, E., Chen, H., 2001. Bringing order to the web: optimizing search by showing results in context. In: Proceedings of ACM Conference on Human Factors in Computer Systems. ACM Press, New York, NY, pp. 277–283 *.

Eisler, H., 1976. Experiments on subjective duration 1868–1975: a collection of power function experiments. Psychological Bulletin 83 (6), 1154–1171.

Fang, X., Salvendy, G., 2000. Keyword comparison: a user-centered feature for improving web search tools. International Journal of Human–Computer Studies 52, 915–931 *.

Farnham, S., Zaner, M., Cheng, L., 2000. Designing for sociability in shared browsers. In: Proceedings of IFIP TC.13 International Conference on Human–Computer Interaction. IOS Press, Amsterdam, pp. 115–123 *.

Farnham, S., Cheng, L., Stone, L., Zaner-Godey, M., Hibbeln, C., Syrjala, K., Clark, A.M., Abrams, J., 2002. HutchWorld: clinical study of computer-mediated social support for cancer patients and their caregivers. In: Proceedings of ACM Conference on Human Factors in Computer Systems. ACM Press, New York, NY, pp. 375–382 *.

Fleiss, J.L., 1981. Statistical Methods for Rates and Proportions. Wiley, New York, NY.

Frøkjær, E., Hertzum, M., Hornbæk, K., 2000. Measuring usability: are effectiveness, efficiency, and satisfaction really correlated. In: Proceedings of ACM Conference on Human Factors in Computer Systems. ACM Press, New York, NY, pp. 345–352.

Garau, M., Slater, M., Bee, S., Sasse, M.A., 2001. The impact of eye gaze on communication using humanoid avatars. In: Proceedings of ACM Conference on Human Factors in Computer Systems. ACM Press, New York, NY, pp. 309–316 *.

Golightly, D., Hone, K.S., Ritter, F.E., 1999. Speech interaction can support problem solving. In: Proceedings of IFIP TC.13 International Conference on Human–Computer Interaction. IOS Press, Amsterdam, pp. 149–155 *.

Gong, L., Lai, J., 2001. Shall we mix synthetic speech and human speech?: impact on users' performance, perception, and attitude. In: Proceedings of CHI2001 ACM Conference on Human Factors in Computer Systems. ACM Press, New York, NY, pp. 158–166 *.

Gould, J.D., Lewis, C., 1985. Design for usability: key principles and what designers think. Communications of the ACM 28 (3), 300–311.

Gutwin, C., 2002. Improved focus targeting in interactive fisheye views. In: Proceedings of ACM Conference on Human Factors in Computing Systems. ACM Press, New York, NY, pp. 267–274 *.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX: results of empirial and theoretical research. In: Hancock, P.A., Meshkati, P. (Eds.), Human Mental Workload. Elsevier, Amsterdam, pp. 139–183.

Hassenzahl, M., Platz, A., Burmester, M., Lehner, K., 2000. Hedonic and ergonomic quality aspects determine a software's appeal. In: Proceedings of ACM Conference on Human Factors in Computing Systems. ACM Press, New York, NY, pp. 201–208.

Hoffman, D.L., Novak, T.P., 1996. Marketing in hypermedia computer-mediated environments: conceptual foundations. Journal of Marketing 60, 50–68.

Hornbæk, K., Frøkjær, E., 2001. Reading electronic documents: the usability of linear, Fisheye, and overview + detail interfaces. In: Proceedings of ACM Conference on Human Factors in Computing Systems. ACM Press, New York, NY, pp. 293–300 *.

Inkpen, K., 2001. Drag-and-drop versus point-and-click mouse interaction styles for children. ACM Transactions on Human–Computer Interaction 8 (1), 1–33 *.

Isbister, K., Naas, C., 2000. Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. International Journal of Human–Computer Studies 53, 251–267 *.

ISO, 1998. Ergonomic requirements for office work with visual display terminals (VDTs)-Part 11: guidance on usability—Part 11: guidance on usability (ISO 9241-11:1998).

Isokoski, P., Käki, M., 2002. Comparison of two touchpad-based methods for numeric entry. In: Proceedings of ACM Conference on Human Factors in Computing Systems. ACM Press, New York, NY, pp. 25–32 *.

Izsó, L., Láng, E., 2000. Heart period variability as mental effort monitor in human computer interaction. Behaviour and Information Technology 19 (4), 297–306 *.

Karat, C.-M., Pinhanez, C., Karat, J., Arora, R., Vergo, J., 2001. Less clicking, more watching: results of the interactive design and evaluation of entertaining web experiences. In: Proceedings of IFIP TC.13 International Conference on Human–Computer Interaction. IOS Press, Amsterdam, pp. 455–463 *.

Kirakowski, J., Corbett, M., 1993. SUMI: the software usability measurement inventory. British Journal of Educational Technology 24 (3), 210–212.

Kurosu, M., Kashimura, K., 1995. Determinants of apparent usability. IEEE International Conference on Systems, Man and Cybernetics, 1509–1513.

LeeTiernan, S., Grudin, J., 2001. Fostering engagement in asynchronous learning through collaborative multimedia annotation. In: Proceedings of IFIP TC.13 International Conference on Human–Computer Interaction. IOS Press, Amsterdam, pp. 472–479 *.

Lewis, C., 1982. Using the "thinking-aloud" method in cognitive interface design. Research Report RC9265, IBM T.J. Watson Research Center.

MacKenzie, I.S., Kauppinen, T., Silfverberg, M., 2001. Accuracy measures for evaluating computer pointing devices. In: Proceedings of ACM Conference on Human Factors in Computing Systems. ACM Press, New York, NY, pp. 9–16 *.

Mankoff, J., Dey, A.K., Hsieh, G., Kientz, J., Ames, M., Lederer, S., 2003. Heuristic evaluation of ambient displays. In: Proceedings of ACM Conference on Human Factors in Computing Systems. ACM Press, New York, NY, pp. 169–176.

Marshall, D., Foster, J.C., Jack, M.A., 2001. User performance and attitude towards schemes for alphanumeric data entry using restricted input devices. Behaviour & Information Technology 20 (3), 167–188 *.

Matarazzo, G., Sellen, A., 2000. The value of video in work at a distance: addition or distraction? Behaviour & Information Technology 19 (5), 339–348 *.

McFarlane, D.C., 1999. Coordinating the interruption of people in human–computer interaction. In: Proceedings of IFIP TC.13 International Conference on Human–Computer Interaction. IOS Press, Amsterdam, pp. 295–303 *.

McGrenere, J., Baecker, R.M., Booth, K.S., 2002. An evaluation of a multiple interface design solution for bloated software. In: Proceedings of ACM Conference on Human Factors in Computing Systems. ACM Press, New York, NY, pp. 163–170 *.

Meister, D., 1985. Behavioral Analysis and Measurement Methods. Wiley, New York.

Miller, R.B., 1971. Human ease of use criteria and their tradeoffs. IBM Technical Report TR 00.2185. IBM Corporation, Poughkeepsie, NY.

Mitsopoulos, E.N., Edwards, A.D.N., 1999. A principled design methodology for auditory interaction. In: Proceedings of IFIP TC.13 International Conference on Human–Computer Interaction. IOS Press, Amsterdam, pp. 263–271 *.

Molich, R., Nielsen, J., 1990. Improving a human–computer dialogue. Communications of the ACM 33 (3), 338–348.

Monk, A., 2002. Noddy's guide to usability. Interfaces 50, 31–33.

Muckler, F.A., Seven, S.A., 1992. Selecting performance measures: "objective" versus "subjective" measurement. Human Factors 34 (4), 441–455.

Mullins, P.M., Treu, S., 1991. Measurement of stress to gauge user satisfaction with features of the computer interface. Behaviour & Information Technology 10 (4), 325–343.

Naur, P., 1965. The place of programming in a world of problems, tools, and people. Proceedings IFIP Congress 65. Also in Peter Naur Computing a Human Activity, 1985. ACM Press, pp. 195–199.

Newman, W., Taylor, A., 1999. Towards a methodology employing critical parameters to deliver performance improvements in interactive systems. In: Proceedings of IFIP TC.13 International Conference on Human–Computer Interaction. IOS Press, Amsterdam, pp. 605–612.

Nichols, S., Haldane, C., Wilson, J.R., 2000. Measurement of presence and its consequences in virtual environments. International Journal of Human–Computer Studies 52, 471–491 *.

Nielsen, J., 1993. Usability Engineering. Academic Press, San Diego, CA.

Nielsen, J., Levy, J., 1994. Measuring usability: preference vs. performance. Communications of the ACM 37 (4), 66–75.

Nyberg, M., Björk, S., Goldstein, M., Redström, J., 2001. Handheld applications design: merging information appliances without affecting usability. In: Proceedings of IFIP TC.13 Conference on Human–Computer Interaction. IOS Press, Amsterdam, pp. 391–398 *.

O'Keefe, R.M., Cole, M., Chau, P.Y.K., Montoya-Weiss, A., Perry, M., 2000. From the user interface to the consumer interface: results from a global experiment. International Journal of Human–Computer Studies 53, 611–628 *.

Pacey, M., MacGregor, C., 2001. Auditory cues for monitoring a background process: a comparative evaluation. In: Proceedings of IFIP TC.13 International Conference on Human–Computer Interaction. IOS Press, Amsterdam, pp. 174–181 *.

Partala, T., 1999. Controlling a single 3D object: viewpoint metaphors, speed and subjective satisfaction. In: Proceedings of IFIP TC.13 International Conference on Human–Computer Interaction. IOS Press, Amsterdam, pp. 486–493 *.

Preece, J., 2000. Online Communities: Designing Usability and Supporting Sociability. Wiley, Chichester, UK.

Raskin, J., 2000. The Humane Interface: New Directions for Designing Interactive Systems. Addison-Wesley, Reading, MA.

Robertson, G.G., Cameron, K., Czerwinski, M., Robbins, D., 2002. Polyarchy visualization: visualizing multiple intersecting hierarchies. In: Proceedings of ACM Conference on Human Factors in Computing Systems. ACM Press, New York, NY, pp. 423–430 *.

Rodden, K., Basalaj, W., Sinclair, D., Wood, K., 2001. Does organisation by similarity assist image browsing? In: Proceedings of ACM Conference on Human Factors in Computing Systems. ACM Press, New York, NY, pp. 190–197 *.

Rosenthal, R., Rosnow, R.L., 1991. Essentials of Behavioral Research: Methods and Data Analysis. McGraw-Hill, Boston, MA.

Rui, Y., Gupta, A., Cadiz, J., 2001. Viewing meeting captured by an omni-directional camera. In: Proceedings of ACM Conference on Human Factors in Computing Systems. ACM Press, New York, NY, pp. 450–457 *.

Sällnas, E.-L., Rassmus-Gröhn, K., Sjöström, C., 2001. Presence in collaborative environments by haptic force feedback. ACM Transactions on Human–Computer Interaction 7 (4), 461–476 *.

Sawnhey, N., Schmandt, C., 2000. Nomadic radio: speech & audio interaction for contextual messaging in nomadic environments. ACM Transactions on Computer–Human Interaction 7 (3), 353–383 *.

Shackel, B., 1959. Ergonomics for a computer. Design 120, 36–39.

Shackel, B., 1981. The concept of usability. Proceedings of IBM Software and Information Usability Symposium. IBM Corporation, Poughkeepsie, NY, pp. 1–30.

Shackel, B., 1991. Usability—context, framework, definition, design and evaluation. In: Shackel, B., Richardson, S. (Eds.), Human Factors for Informatics Usability. Cambridge University Press, Cambridge, pp. 21–38.

Shneiderman, B., 1998. Designing the User Interface. Addison-Wesley, Reading, MA.

Shneiderman, B., 2000. Creating creativity: user interfaces for supporting innovation. ACM Transactions on Computer–Human Interaction 7 (1), 114–138.

Slater, M., Sadagic, A., Usoh, M., Schroeder, R., 2000. Small group behaviour in a virtual and real environment. Presence: Teleoperators And Virtual Environments 9 (1), 37–51.

Smith, S.L., Mosier, J.N., 1986. Guidelines for designing user interface software. Mitre Report ESD-TR-86-278, The MITRE Cooperation, Bedford, MA.

Soloway, E., Guzdial, M., Hay, K.E., 1994. Learner-centered design, Interactions, April, 36–48.

Speilberger, C.D., 1983. Manual for the State-Trait Anxiety Inventory (STAI). Consulting Psychologists Press, Palo Alta, CA.

Sweeney, M., Maguire, M., Shackel, B., 1993. Evaluating user-computer interaction: a framework. International Journal of Man–Machine Studies 38 (4), 689–711.

Tan, D.E., Robertson, G.G., Czerwinski, M., 2001. Exploring 3D navigation: combining speed-coupled flying with orbiting. In: Proceedings of ACM Conference on Human Factors in Computing Systems. ACM Press, New York, p. NY *.

Tattersall, A.J., Foord, P.S., 1996. An experimental evaluation of instantaneous self assessment as a measure of workload. Ergonomics 39, 740–748.

Tractinsky, N., 1997. Aesthetics and apparent usability: empirically assessing cultural and methodological issues. In: Proceedings of ACM Conference on Human Factors in Computing Systems. ACM Press, New York, NY, pp. 115–122.

Tractinsky, N., Meyer, J., 2001. Task structure and the apparent duration of hierarchical search. International Journal of Human–Computer Studies 55, 845–860 *.

Tyfa, D.A., Howes, M., 2000. Speech recognition for command entry in multimodal interaction. International Journal of Human–Computer Studies 52, 637–667 *.

Underhill, P., 2000. Why We Buy: The Science of Shopping. Touchstone Books.

Wang, Q.Y., Shen, M.W., Shui, R., Su, H., 2001. Detectability and comprehensibility study on audio hyperlinking methods. In: Proceedings of IFIP TC.13 International Conference on Human–Computer Interaction. IOS Press, Amsterdam, pp. 310–317 *.

Wastall, D., 1990. Mental effort and task performance: towards a psychophysiology of human computer interaction. In: Proceedings of Third

IFIP International Conference on Human Computer Interaction. North-Holland, Amsterdam, pp. 107–112.

Wheeless, L.R., Grotz, J., 1977. The measurement of trust and its relationship to self-disclosure. Human Communication Research 3 (3), 250–257.

Westerman, S.J., Cribbin, T., 2000. Mapping semantic information in virtual space: dimensions, variance and individual differences. International Journal of Human–Computer Studies 53, 765–787.

Westerman, S., Cribbin, T., Wilson, R., 2001. Virtual information space navigation: evaluating the use of head tracking. Behaviour and Information Technology 20 (6), 419–426 *.

Wharton, C., Rieman, J., Lewis, C., Polson, P., 1994. The cognitive walkthrough method: a practitioner's guide. In: Nielsen, J., Mack, R.L. (Eds.), Usability Inspection Methods. Wiley, New York, pp. 105–140.

Whiteside, J., Bennett, J., Holtzblatt, K., 1988. Usability engineering: our experience and evolution. In: Helander, M. (Ed.), Handbook of Human–Computer Interaction. Elsevier, Amsterdam, pp. 791–817.

Whittaker, S., Hircshberg, J., Amento, B., Stark, L., Bacchiani, M., Isenhour, P., Stead, L., Zamchick, G., Rosenberg, A., 2002. SCANMail: a voicemail interface that makes speech browsable, readable and searchable. In: Proceedings of ACM Conference on Human Factors in Computing. ACM Press, New York, NY, pp. 275–282 *.

Woodruff, A., Faulring, A., Rosenholtz, R., Morrison, J., Pirolli, P., 2001. Using thumbnails to search the web. In: Proceedings of ACM Conference on Human Factors in Computing. ACM Press, New York, NY, pp. 198–205 *.

Wulf, V., Golombek, B., 2000. Direct activation: a concept to encourage tailoring activities. Behaviour and Information Technology 20 (4), 249–263 *.

Yeh, Y.-Y., Wickens, C.D., 1988. Dissociation of performance and subjective measures of workload. Human Factors 30 (1), 111–120.

Zhai, S., 2004. Characterizing computer input with Fitts' law parameters—the information and non-information aspects of pointing. International Journal of Human–Computer Studies 61, 791–809.