Computers in Human Behavior 54 (2016) 101–113

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Metacognition and system usability: Incorporating metacognitive research paradigm into usability testing ☆

Rakefet Ackerman *, Avi Parush, Fareda Nassar, Avraham Shtub

Faculty of Industrial Engineering and Management, Technion–Israel Institute of Technology, Haifa, Israel

ARTICLE INFO

Article history:
Received 30 August 2014
Received in revised form
19 July 2015
Accepted 20 July 2015
Available online 11 August 2015

Keywords:
Usability testing
Summative testing
Perceived usability
Cognitive processes
Metacognition
Confidence

ABSTRACT

There is an agreement that perceived usability is important beyond actual effectiveness of software systems. Perceived usability is often obtained by self-reports provided after system use. Aiming to improve summative usability testing, we propose a methodology to enhance in-depth testing of users' performance and perceived usability at the task level. The metacognitive research approach allows detailed analysis of cognitive processes. Adapting its methodologies, we propose the Metacognitive Usability Profile (MUP) which includes a comprehensive set of measures based on collecting confidence in the success of each particular task and triangulating it with objective measures. We demonstrate using the MUP by comparing two versions of a project management system. Based on a task analysis we allocated tasks that differ between the versions and let participants ($N = 100$) use both versions. Although no difference was found between the versions in system-level perceived usability, the detailed task-level analysis exposed many differences. In particular, overconfidence was associated with low performance, which suggests that user interfaces better avoid illusions of knowing. Overall, the study demonstrates how the MUP exposes challenges users face. This, in turn, allows choosing the better task implementation among the examined options and to focus attempts for usability improvement.

1. Introduction

An important part of the quality assessment of software systems, is assessing their usability. According to Nielsen (1993), usability consists of learnability, efficiency, memorability, low error rate or easy error recovery, and satisfaction. This conceptualization combines objective and subjective aspects of success in using the examined system (see Albert & Tullis, 2013; for a review). The present study contributes to utilizing this combination in usability tests by adapting methods from the metacognitive research domain. This domain combines objective and subjective measures of cognitive performance in contexts such as learning and problem solving.

Perceived usability is central in the subjective aspect of usability and can influence users' decision regarding purchase and extent of system's use (see Hertzum, 2010 for a review). Self-reports are generally considered as good approximations of perceived usability which can be obtained, among others, by having users respond to questionnaires after interacting with the system. Post-interaction self-reports can be system-oriented and/or task-specific. A very popular perceived usability scale which is system-oriented is the ten-item System Usability Scale (SUS; Brooke, 1996). Its strength is in providing a single score that allows comparing perceived usability across diverse systems (see Bangor, Kortum, & Miller, 2008 for a review). While the SUS is still much in use, there have been further developments in the evaluation of perceived usability (J. R. Lewis, 2015a, b). Those include primarily questionnaires that are shorter such as the Usability Metric for User Experience (Bosley, 2013; Finstad, 2010; J. R. Lewis, Utesch, & Maher, 2015) and questionnaires that consider the emotional and experiential aspects of usability, such as the Emotional Metric Outcome Questionnaire (Borsci, Federici, Bacci, Gnaldi, & Bartolucci, 2015; J. R. Lewis & Mayes, 2014). Nevertheless, since most developments in assessing system-oriented subjective usability relate to the SUS, we used it in this study as a benchmark reflecting system-oriented perceived usability.

☆ This research was partially supported by the Israeli Science Foundation research grant number 563/09: Designing Visualization for Expert Training in Project Management.

* Corresponding author. Faculty of Industrial Engineering and Management, Technion, Technion City, Haifa, 3200003, Israel.
E-mail address: ackerman@ie.technion.ac.il (R. Ackerman).

http://dx.doi.org/10.1016/j.chb.2015.07.041
0747-5632/© 2015 Elsevier Ltd. All rights reserved.

Two challenges are readily apparent with such system-oriented self-reports. First, retrospective assessments of usability may be biased toward the most recent (or salient) experiences and, thus, not be representative of the overall experience with the system (Hassenzahl & Sandweg, 2004). Second, when used as part of summative usability testing, they do not necessarily reflect usability at the task level. For example, Callan and Kortum (2014) showed that task-specific SUS scores were significantly higher than system-oriented SUS scores for the same system. As a result, system-oriented perceived usability can be limited in guiding designers in improving usability of a given system. For focusing improvement attempts on task-level issues, a more fine-tuned methodology is required.

Summative usability testing typically takes place when the product design and development are complete or close to completion. Yet, there is still a need for the summative testing to support procurement decisions, guide the development of help and support systems, and to guide revisions and updates of the product. For these purposes, there is a need for in-depth information regarding the nature and source of the usability problems in much more detailed level (Høegh & Jensen, 2008).

Task-specific subjective measures, such as the After Scenario Questionnaire for Computer Usability (J. R. Lewis, 1991) or One-question Post-task questionnaires (Sauro & Dumas, 2009), provide focused usability scores (Sauro & Dumas, 2009; Sauro & Lewis, 2009). However, correlations between task-specific objective measures, such as response time, and subjective measures, such as post-task or post-test satisfaction, are often low or inconsistent (Hornbæk & Law, 2007).

Methodologies from cognitive psychology research have been adapted to the usability testing domain aimed at extracting detailed task-specific information. These methodologies are relevant because while interacting with a system, people read, reason, solve problems, etc., which are all complex cognitive processes studied extensively. Of particular relevance for combining objective and subjective usability measures are the cognitive think aloud and walkthrough techniques which are often used for inferring system's usability based on user's cognitive processes during its use (e.g., Nielsen, 1993).

Think aloud is a well-known technique in cognitive psychology in which the participants are asked to verbalize their ongoing thinking during task performance (see Ericsson & Simon, 1993). This method is used in the context of usability tests to uncover thinking processes and subjective experience that cannot be examined behaviorally (e.g., McDonald, Zhao, & Edwards, 2013). Cognitive walkthrough is a step-by-step process whereby users evaluate aspects related to perceived usability and understanding while using the system (C. Lewis, Polson, Wharton, & Rieman, 1990). However, empirical studies that examined the effect of thinking aloud and cognitive walkthrough on the outcomes of usability tests imply that those techniques can be disruptive to task performance as reflected by objective measures (e.g., Hertzum, Hansen, & Andersen, 2009; Hertzum & Holmegaard, 2013; see J. R. Lewis, 2014; for a review).

Taken together, findings of research so far imply there is still a need for a methodology that focuses on exposing the cognitive processes involved while using systems, and yet can reflect both objective and subjective aspects at the task level. Applying the metacognitive paradigm to usability testing can provide the benefits of using a non-disruptive technique and yet be task-specific. As outlined below, recent human–computer interaction studies have analyzed metacognitive aspects of learning texts on screen versus on paper and provided insights regarding media effects on cognitive processing (e.g., Lauterman & Ackerman, 2014). In the present study, we extend this work by adapting the metacognitive paradigm as an innovative task-specific usability scorecard that taps into the cognitive processes taking place during interaction with software tools. We suggest that this detailed analysis can support choice among implementation variations and guide further design iterations even in summative testing.

1.1. Metacognition

The metacognitive approach puts forward the centrality of subjective judgment of confidence in the success of performing cognitive tasks (see Bjork, Dunlosky, & Kornell, 2013; for a review). The importance of the extent people feel confident in performing tasks is well acknowledged in the usability literature. Some studies refer to users' confidence by inferring it from self-report of assessed success level (e.g., Cassidy, Jones, McMain, Shen, & Vieira, 2014), while others ask about confidence explicitly (e.g., Jones et al., 2008). However, the metacognitive approach goes beyond comparing confidence levels among conditions or people, by analysis of relationships between confidence, actual success rate, and response time across several tasks. This is because a central notion in this literature is that people use their judgment regarding each subtask (e.g., question) for deciding whether to invest more time or move on to the next subtask. Unreliable judgments mislead the investment of time and effort, which, in turn, could degrade performance (e.g., Metcalfe & Finn, 2008; Thiede, Anderson, & Therriault, 2003).

There are two common measures for judgment accuracy, calibration and resolution, which tap orthogonal aspects of it. *Calibration bias* (or absolute accuracy) reflects deviations of individuals' judgments from their actual chance for success when averaging several task items. For example, when participants answer a set of questions that each of them is accompanied by a confidence rating on a 0–100% scale, a positive bias score represents overconfidence (e.g., average confidence of 80% with average success rate of 70%), and a negative bias score represents under-confidence. Overall, people tend to be overconfident (Dunning, Heath, & Suls, 2004; Metcalfe, 1998). Overconfidence is problematic (Dunlosky & Rawson, 2012) because when users think that they perform the task adequately while in fact they perform it poorly, this illusion will prevent them from acting to improve their performance (e.g., open the user manual). Of course, having frustrated users, who have low confidence in their ability to succeed in performing tasks with the system, is not desirable as well. Thus, we would like to stress that user interfaces (UIs) which facilitate reliable confidence across the various tasks have an advantage over those leading to illusion of success, despite the immediate satisfaction that may come with it.

Resolution (or relative accuracy) reflects the extent in which judgments discriminate between successful and unsuccessful tasks. It is measured by correlating judgments and success across tasks within participant (e.g., Metcalfe & Finn, 2008). Perfect resolution (correlation of 1 within the range -1 to $+1$) is achieved when higher judgments are assigned for all the successful tasks than to the tasks in which the participant was less successful. One can be highly overconfident, but still have perfect resolution, and vice versa. For example, let us assume that a given participant provided 90% confidence rating whenever performing a task correctly, and 85% for all wrong responses. This participant shows perfect resolution, since confidence ratings discriminate perfectly between the correct and wrong responses. However, if the actual success rate over the entire task was 60%, then this participant showed a pronounced calibration bias, in the form of overconfidence.

Another aspect of interest is the association between response time, on the one hand, and actual chance for success and subjective confidence, on the other. In the context of metacognitive theory, this analysis is used for studying the underlying heuristic cues that