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

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.
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 (Hornbak & 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, & Theriault, 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; Mcalpage, 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
inform confidence judgments (e.g., Ackerman & Koriat, 2011; Kelley & Lindsay, 1993) and the stopping rule people adapt for their effort investment (Ackerman, 2014). In the context of eyewitness testimony, when asked a question to be answered based on memory for event’s details, the combination of response time and confidence is used to predict the chance of an information piece to be correct (e.g., Sauerland, Sagana, & Sporer, 2012). The common finding in these bodies of literature is a negative correlation between response time and both confidence and actual chance for finding the correct answer. This is because the easy questions are answered quickly with high confidence and high chance for success, while the more challenging questions still have lower chance for success even after lengthy thinking, and respondents acknowledge it (Ackerman, 2014; Koriat, Ma’ayan, & Nussinov, 2006).

Possible reasons for quick responses in task performance may be (a) the task is simple and requires only a few quick steps (b) the task is perceived as easy while it is in fact more complicated, or (c) the participant has no clue what to do and how to address the task and thus gives it up quickly. We suggest that associating response time with success rate and confidence in the context of usability tests may introduce a unique contribution. We would expect to find high confidence which is reliable in case a, overconfidence in case b, and low and reliable confidence in case c. On the other hand, people may invest a lot of effort trying to perform a task which they do not find how to perform. After lengthy thinking, people tend to give up and move on to the next task, when possible (Ackerman, 2014).

Confidence ratings allow also the distinction between slow responses which stem from failure to find how to address the question while acknowledging it, and an effortful thought course which the participant believes to be successful (reliably or not).

1.2. Metacognitive Usability Profile (MUP)

We propose the Metacognitive Usability Profile (MUP) as a methodology for in-depth comparison of two or more UIs intended for performing similar detailed tasks. We see the MUP methodology as particularly appropriate when analyzing complex systems that require high-order cognitive processing to be performed accurately and efficiently. Importantly, like the tasks used in the metacognitive research, the examined tasks should be well-defined so to allow assessing performance, rather than being flexible (e.g., graphical design) and/or involve opinion (e.g., beauty assessment). However, success should not be readily apparent (e.g., having the entire picture in a jigsaw puzzle versus having several pieces for which the proper place could not be found). This ambiguity in task success is required for allowing participants to assess their own chance for success and express it in terms of confidence ratings (0–100%) with variability (not only 0 or 100% like in the jigsaw puzzle example). The experimenters may assess objective success in each task either in terms of dichotomy (0-failure or 100-success) or by a continuous percentage-based (0–100%) grade. This symmetry between the subjective self-assessment of confidence and objective assessment of performance allows associating them in each particular task and across tasks.

In the present study, we demonstrate using the MUP methodology and its beneficial outcomes by comparing two UIs of a project management tool (see details below). In this example, project management could be broken into tasks such as hiring and firing workers, using a particular resource (e.g., building materials), or assessing the monetary balance in a midpoint of the project. Such a system satisfies all the above detailed criteria for fitting the MUP.

For analyzing the usability of the selected tasks and comparing it between the UIs, the MUP includes traditional usability measures, such as performance, response time, and perceived task difficulty, as well as novel measures for this context, as detailed below. The combined measure set allows the derivation of a usability scorecard (e.g., Albert & Tullis, 2013) which provides a comprehensive task-specific picture of the differences between the UIs, pointing to the relative strengths and weaknesses of each one of them.

The measures that comprise the MUP were derived from the Metacognitive Learning Regulation Profile (MLRP) suggested by Ackerman and Goldsmith (2011) for comparing learning environments. Ackerman and Goldsmith compared learning texts presented on the computer screen to learning the same texts from paper by comparing a set of measures comprised of direct (e.g., success rate) and indirect (calculated) measures (e.g., overconfidence). Importantly, unlike previous studies which attributed screen inferiority in learning outcomes to technological disadvantages (e.g., navigation within the document, markup tools, screen glare), they found that learning efficiency per se did not differ between the media, while the metacognitive processes were inferior on screen. The participants were consistently more overconfident on screen than on paper, and regulated their efforts less effectively (see also Ackerman & Lauterman, 2012; Lauterman & Ackerman, 2014).

The MUP adapts from the MLRP the direct measures of success rate, response time, and confidence (replaces predicting learning outcomes after test). It also adapts the indirect measures of efficiency (points gained per minute of work), calibration bias (overconfidence), and resolution. On top of these, we added to the MUP the perceived task difficulty, which is particularly relevant for usability tests. We also included in the MUP time-success and time-confidence relationships, which reflect the recent development in metacognitive research which differentiates quick and slow responses, as explained above.

In summary, considering usability under the umbrella of the metacognitive paradigm could provide UI designers with considerable added value in assessing task performance and cognitive aspects and their relation to perceived usability. In particular, we suggest that the MUP allows exposing specific usability problems beyond what is exposed by system-oriented questionnaires such as the SUS. This in turn can facilitate allocating candidate tasks for usability improvement by adding unique information to the currently known task-specific methods.

1.3. The present study

In the present study, we used the SUS as well as the proposed MUP’s comprehensive set of measures for comparing the usability of two versions of a project management system. These two versions have exactly the same functionality and share the same server program. Thus, any difference that may be found should be attributed to the UI design.

We started the analysis by choosing the set of tasks to be examined in both UI versions. We focused on tasks done during ordinary project management and applied task analysis for identifying tasks that are likely to expose differences between the versions. This was done by watching and interviewing participants sampled from the target population working with both UI versions. The collected information guided the choice of nine focused tasks divided into two task sets; each of them was expected to be easier in one version compared with the other.

Each task was a basis for a question participants had to answer by performing this task in the system (see example below). The examination was done by collecting confidence in each answer, in addition to success, response time, and perceived task difficulty. After performing the nine tasks, the 10th task was to perform a whole project management run. Participants rated their confidence and perceived task difficulty for the whole run, as for other tasks. Finally, immediately after performing the whole project
run, participants rated their global experience with the UI by the SUS.

Previous studies pointed to differences in usability generated by users’ knowledgability in the target domain (e.g., Sauer, Seibel, & Rüttinger, 2010). Thus, in the present study, the participants were students before and after taking the project management course, although none of them had experience with actual project management nor with project management systems.

The MUP was used to examine which UI version and task set generated better and more efficient task performance and more reliable confidence judgments. In addition, we compared the outcome of SUS to the results of the analysis by MUP.

2. Method

2.1. Participants

One hundred Engineering students participated in the study (45 females; M_age = 26.1, SD = 2.7; 87 for course credit and the rest for payment). Forty nine participants graduated the Project Management course at the Faculty of Industrial Engineering and Management of the Technion—Israel Institute of Technology.

2.2. Materials

2.2.1. The compared user interfaces

Project management is a complicated engineering area that is characterized by requiring taking into account large variety of constraints (Davidovitch, Parush, & Shtub, 2006). The two UI versions were two front sides of one project-management system, Project Team Builder (PTB; see Shtub, 2013). They were designed to represent the various project components in a visual-spatial form, making the information easily accessible. The PTB is used to provide students in an academic Project Management course with hands-on experience of managing projects in a dynamic stochastic environment, which takes into account unexpected events, like absence of a worker with critical skills.

The input of the PTB is a project scenario, which includes tasks that compose the project with precedence relations among them, cost, target date, and the number of workers required in each period and their expertise. Different scenarios describe different projects. Users’ aim is to plan the project and to execute it in the shortest time and lowest budget possible.

The main difference between the two UI versions is that in one version, PTB1, all the information is presented simultaneously on the screen (see Fig. 1). This allows the toolbar to be short (see Fig. 2). In the other version, PTB2, the information regarding the various project components (tasks, costs, and resources) is presented in separated panels. For example, Fig. 3 presents the tasks in the project. In order to view cost or resources, a dedicated panel should be opened. In the toolbar, the commands and panel activation buttons appear side by side (see Fig. 4). This version allows spacious presentation of each information type, but requires the users to integrate the information in their memory.

2.2.2. Experimental materials

Besides the two UI versions, there were instructions regarding the experiment and regarding basic project management terminology common to both versions. Participants’ work was guided by a form which included a page for filling in personal details and the task questions. Each question represented a task chosen by the task analysis and up to two questions appeared on each page. Each question appeared with a space for an answer, a scale for rating confidence (0—100%), and a Likert scale for rating perceived task difficulty (1 — Very easy and 7 — Very difficult). See a question example with a demonstration how the scales were used in Fig. 5.

The nine tasks included Set1 (4 tasks), expected to be easier in PTB1, and Set2 (5 tasks), expected to be easier in PTB2. The tasks of the two sets were mixed, with their order determined by real-life flow of implementing a project simulation run (e.g., firing a worker is possible only after recruitment). The order of the tasks in the questionnaire is detailed in the Appendix, with the reason for expecting a particular challenge in either of the two UI versions.

Unlike the first nine tasks, which involved a particular step in the cycle of project management, the 10th task involved a continuous project management based on a given project scenario. It involved performing a task-sequence similar to real-life project management. The participants had control over project’s duration and cost by hiring and firing personnel while planning the resources and taking into account the stochastic characteristics of the scenario (e.g., sick leave of a critical worker or an unplanned delay in performing a task). Performance in this task was measured by the remaining cash at the end of the project (the larger is better) and the estimated duration of the project (the shorter the better).

2.3. Procedure

The experiment was conducted with each participant separately. At the beginning of the session, all pages were placed upside down next to the participant, pre-arranged according to their order of use throughout the experiment. With each step, the participant turned over the relevant page. The participant first read the instructions booklet, and then performed the nine specific tasks and the project simulation with the first version and first project scenario. Response times were captured by the experimenter (in seconds) in the above mentioned table, by using a stop watch. The participant then filled in the SUS regarding the first version. Immediately after that, the participant moved on to the second version, for which the entire procedure repeated with a different scenario. Counterbalancing took place allowing all four possible pairing between the two UI versions and the two scenarios across participants. The procedure did not allow skipping a question or getting back to a previous one. In a case of failure to perform a task, the participant was encouraged to answer the question to the best of his/her knowledge, and convey the perceived low chance for success by the confidence scale. A session took about two hours per participant.

3. Results

3.1. Control variables

Compensation type and the scenario used for the project simulation had no effect on the results. As well-acknowledged in literature, despite no difference in success rates, t = 1, males provided higher confidence ratings (M = 83.2, SD = 11.0) than females (M = 74.3, SD = 15.4), t(98) = 3.40, p = .001, were more over-confident (M_males = 10.4, SD = 12.9; M_females = 2.3, SD = 14.1), t(98) = 3.07, p = .003, and provided marginally lower perceived task difficulty ratings (M_males = 2.9, SD = 0.8; M_females = 3.2, SD = 0.8), t(98) = 1.96, p = .053. Importantly, there were no interaction effects between gender and the UI version. Thus, the analyses reported below were conducted across genders.

Notably, no difference was found in their SUS scores between participants who graduated a project management course
Fig. 1. Task management window of PTB1.

Fig. 2. PTB1 toolbar.

Fig. 3. Task management window of PTB2.
(knowledgeable) and those who did not take the course yet (novices). In addition, there was no interaction effect with the version, both \( F_s < 1 \). Thus, the two UI versions were perceived by the participants sub-groups as equally usable (according to the SUS score), regardless of their familiarity with project management concepts.

Nevertheless, the detailed analysis revealed that the more knowledgeable participants had higher success rates relative to novices (\( M_{\text{knowledgeable}} = 76.1, SD = 12.7; M_{\text{novice}} = 69.0, SD = 15.8 \), \( t(98) = 2.45, p = .016 \). They were also more confident in their answers (\( M_{\text{knowledgeable}} = 77.7, SD = 20.5; M_{\text{novice}} = 96.7, SD = 26.4 \), \( t(98) = 4.02, p < .0001 \), and were faster to respond (\( M_{\text{knowledgeable}} = 77.3, SD = 20.5; M_{\text{novice}} = 96.7, SD = 26.4 \), \( t(98) = 4.07, p < .0001 \). As a result, the knowledgeable participants were more efficient (\( M_{\text{knowledgeable}} = 7.0, SD = 2.4; M_{\text{novice}} = 5.2, SD = 2.1 \), \( t(98) = 4.07, p < .0001 \), but their overconfidence was equivalent to that of the novices, \( t = 1 \). No interaction effects with the version were found the version. Thus, the analyses reported below were conducted across all participants.

### 3.2. Comparing the user interfaces

The SUS and the performance of the continuous project management (10th task), represent the system-level measures often collected and analyzed in usability tests. The means (and SD) and results of Analysis of Variance (ANOVA) comparing the two versions in these measures are presented in Table 1. The two versions were equivalent by the SUS and by most of the measures regarding the continuous project management task. The only difference found was the longer project duration with PTB1 than with PTB2. This finding reflects an advantage for PTB2.

A schematic representation of the MUP regarding performance of the first nine tasks is presented in Fig. 6. The figure provides a comparison between the two UI versions, while the detailed ANOVAs are presented in Table 2. In Fig. 6, the MUP of each UI version is represented by a line, blue with circles for PTB1 and red with squares for PTB2. Y-axis represents the comparison result, such that the higher is better (for response time the scale was reversed). MUP components for which it is clear what values are accounted to be better (e.g., higher success rate, quicker performance) are represented on the X-axis. Other components (time-success rate and time-confidence relationships) are not represented in the figure, but discussed below. Significant differences (\( p < .05 \)) are marked by a large distance between the points, no difference is represented by touching points, and a small distance represents a marginally significant difference, \( p \approx .08 \). Panel A presents the results across the nine tasks. Panel B and Panel C separate the results by task set.

### Table 1

Means (SD) and ANOVA results when comparing the two user interface versions, PTB1 and PTB2, by global measures.

<table>
<thead>
<tr>
<th>Measure</th>
<th>PTB1</th>
<th>PTB2</th>
<th>( F(1.99) )</th>
<th>( \eta^2_p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Usability Scale (SUS)</td>
<td>57.9 (20.8)</td>
<td>56.2 (18.7)</td>
<td>&lt;1</td>
<td>.00</td>
</tr>
<tr>
<td>Continuous project management (10th task): Success rate, cash amount at the end of the project, time to perform the task, confidence, and perceived task difficulty</td>
<td>Most measures</td>
<td>Most measures</td>
<td>&lt;1</td>
<td>.00</td>
</tr>
<tr>
<td>Continuous project management (10th task): Project duration (in weeks)</td>
<td>37.3 (4.1)</td>
<td>36.3 (3.3)</td>
<td>2.76*</td>
<td>.07</td>
</tr>
</tbody>
</table>

* Effect sizes of were measured by \( \eta^2_p \) (partial eta-squared) with .01, .06, and .14 representing small, medium, and large, respectively (based on Cohen, 1988; see Richardson, 2011). ANOVA was used here for allowing comparison of effect sizes with the following analyses, which were in a 2 x 2 design.
The schematic view in Panel A of Fig. 6 suggests that the two versions are equivalent for three measures, PTB1 is better in one measure, PTB2 is better in two measures, and there is one marginal advantage for PTB2. Thus, overall there is some advantage for PTB2, although not a consistent one. The more detailed pictures seen in Panel B and Panel C provide strong support in the task analysis which guided the selection of tasks for Set1 and Set2 for exposing differences between the versions. In fact, differences between the versions and/or interaction effects with the task set were found in all MUP measures.

Table 2 presents the means (and SD) and ANOVA results of the combined effects of Version (PTB1 vs. PTB2) and Set (Set1 vs. Set2) on the various measures. Panel B and Panel C of Fig. 6 summarize the results and represent graphically the MUP per task set. The ANOVA results were the basis for drawing Panel B and Panel C of Fig. 6. The main effect of the set is not reported, because the sets were not designed to be equivalent, and differences, where found, may be misleading. Our focus is on allocation usability problems by pointing to the tasks in which the two versions differ.

We now demonstrate the insights that can be drawn from this detailed analysis. In this analysis, we emphasize the usability implications the MUP components expose.

3.2.1. Overview of MUP components

Panel B and Panel C of Fig. 6 draw attention to the lower line in each panel. These lines point to features, those included in the incompatible set for each version, that call for usability
improvement. Components in which the two lines are equivalent or even switched for a particular set in a particular component (confidence and resolution in Set1 and response time in Set2) should also be analyzed with care, for understanding how come the version that supports this set better in other components loses its advantage in these.

3.2.2. Success rate, response time, and efficiency

The relationship between success rate and response time allows considering efficiency. For this purpose, we referred to success rate as points, very much like in school tests (70% correct is accounted as 70 gained points). Dividing the gained points by the invested minutes generated a measure of efficiency, representing the points gained per minute of work.

Despite the overall equivalence between the versions, the ANOVA which took the task set into account revealed a strong interaction effect on efficiency. While performing Set1 tasks in PTB1 was faster and more successful than in PTB2, when facing Set2, although the success rate was better in PTB2, response time was equivalent between the two versions (see Panel B and Panel C of Fig. 6). This finding points to an advantage for PTB1 over PTB2, and should guide re-design of these features.

The fact that success rates and efficiency were better in the compatible version for each task set support the task analysis that led to the choice in these nine tasks for the purpose of demonstrating using the MUP. It highlights that the MUP is sensitive to the usability differences for which UI designers are sensitive too. That is, it suggests that if these tasks would have been chosen by other criteria (e.g., the most recent development in the system), the MUP would allow exposing usability differences between the versions.

3.2.3. Perceived task difficulty and confidence

One may expect perceived task difficulty and confidence to be two expressions for the same experience. The results show a similar, but not identical, pattern of results. In both measures, PTB2 reflected subjectively easier (or marginally easier) performance of the tasks across the two sets (see Panel A of Fig. 6). Delving into the combined (interaction) effect of version and set on perceived task difficulty and on confidence showed that the difference between the versions in Set1 was significant for perceived task difficulty, while for confidence the ratings were equivalent. For Set2 the two measures showed similar advantage for PTB2. Future studies are called for delving into the meaning participants attach to the two measures, and the differential cues that underlie each of them.

3.2.4. Confidence, success rate, and overconfidence

An important finding that the metacognitive analysis highlights is that differences between the versions in confidence and actual success rates show distinctive patterns. For example, the higher confidence in PTB 2 (depicted as “better” in the MUP) was not associated with an expected corresponding higher success rates (see panel A in Fig. 6). Another example is the equivalence in confidence ratings for the two versions regarding Set1, while performance was in fact better in PTB1.

These discrepancies sometimes result in differences in overconfidence (see Panel B and Panel C, but not in Panel A, of Fig. 6). This depends on the variance of success rates and confidence. The largest overconfidence bias was found with PTB2 for Set1, reflecting that the incompatibility of Set 1 of tasks to PTB 2 (as was expected by our task analysis) was not expressed by the participants. In both versions, the participants were well calibrated (overconfidence was not significantly larger than zero) for the compatible sets and overconfident for the incompatible ones. This is a demonstration of the classic hard-easy effect in the metacognitive literature, by which overconfidence is typically larger for the harder tasks than for the easier ones (Juslin, Winman, & Olsson, 2000).

3.2.5. Resolution – differentiation between correct and wrong responses

The hard-easy effect stems from the smaller differences in confidence ratings between the sets compared with larger differences in success rates. This finding is important because it may reflect users’ difficulty in differentiating between correct and wrong responses. Such a possibility, as will be discussed later, has implications for usability analysis.

As explained above, resolution measures exactly this aspect. Success is a dichotomy—an answer is either correct or wrong—and confidence is a continuous measure (0–100%). Thus, following Nelson (1984), resolution is typically measured by the a-parametric Gamma correlation (see Masson & Rotello, 2009; for a recent discussion about pros and cons of gamma). Calculating such within-participant gamma correlation requires at least six items per participant, and the more is better. In addition, this statistic measure removes from the analysis cases in which there is no variability in either correctness or confidence ratings. In the present study there were only four or five items in each set. Thus, we could not use gamma correlations for analyzing resolution per set.

Comparing the two versions across the sets by gamma correlation for the participants who showed variability in both measures.

Table 2

<table>
<thead>
<tr>
<th></th>
<th>PTB1</th>
<th>PTB2</th>
<th>Version main effect</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Set1</td>
<td>Set2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct measures</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Success rate</td>
<td>75.7 (21.8)</td>
<td>67.5 (23.7)</td>
<td>1.48 .02</td>
<td>53.18*** .35</td>
</tr>
<tr>
<td>Response time (sec.)</td>
<td>100.9 (44.1)</td>
<td>55.9 (31.5)</td>
<td>5.65 .05</td>
<td>23.0*** .19</td>
</tr>
<tr>
<td>Perceived task difficulty</td>
<td>3.1 (1.3)</td>
<td>3.1 (1.3)</td>
<td>3.15 .03</td>
<td>78.44*** .44</td>
</tr>
<tr>
<td>Confidence</td>
<td>79.2 (18.7)</td>
<td>76.5 (19.3)</td>
<td>4.96 .05</td>
<td>64.33*** .39</td>
</tr>
<tr>
<td>Derived measures</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Efficiency (points gained per min.)</td>
<td>10.7 (5.5)</td>
<td>25.1 (18.9)</td>
<td>&lt;1 .00</td>
<td>12.1*** .11</td>
</tr>
<tr>
<td>Overconfidence</td>
<td>3.6 (20.6)</td>
<td>8.9 (24.3)</td>
<td>&lt;1 .00</td>
<td>18.8*** .16</td>
</tr>
<tr>
<td>Resolution</td>
<td>-.47 (.69)</td>
<td>.22 (.86)</td>
<td>7.36*.07</td>
<td>&lt;1 .01</td>
</tr>
<tr>
<td>Time–success relationship</td>
<td>-.34 (.78)</td>
<td>.22 (.84)</td>
<td>.05 (.72)</td>
<td>&lt;1 .01</td>
</tr>
<tr>
<td>Time–confidence relationship</td>
<td>-.65 (.67)</td>
<td>.42 (.90)</td>
<td>.06 (.85)</td>
<td>.35 .03</td>
</tr>
</tbody>
</table>

\* p < .05, ** p < .01, *** p < .001.
\* Significantly larger than zero.
for both versions, revealed marginally weaker resolution in PTB1 ($M = .35$, $SD = .64$) than in PTB2 ($M = .50$, $SD = .54$), $t(82) = 1.77$, $p = .08$. Thus, PTB1 supported somewhat less reliable discrimination between the correct and wrong responses. Importantly, for both versions the resolution was significantly larger than zero, $ps < .0001$. It means that the confidence ratings were higher for the correct answers than for the wrong answers. This resolves the doubt rose above, regarding the sensitivity of confidence ratings to success in the tasks.

Delving into resolution differences at the level of the two task sets was done by approximation to gamma correlation. We split the data of each participant into correct and wrong responses, and calculated the mean confidence for each. In line with the procedure of gamma correlation, a positive confidence difference was marked as fit (1), and a negative difference was marked as unfit ($-1$). In addition, we marked as fit (1) cases in which all answers were correct and all confidence ratings were 100%. Other cases were marked as reflecting no discrimination (0). This measure allows including in the analysis the data of all participants. Resolution means (and SDs) per version and set are presented in Table 2. As can be seen, participants’ resolution was reliable (significantly larger than zero) by this measure as well. Overall, resolution was significantly weaker in PTB1 than in PTB2. Interestingly, while it was significantly weaker for PTB1 when comparing the versions for task Set2, which is in line with the other measures with incompatible task set, it also tended to be weaker in PTB1 when considering its compatible set, Set1, $t(99) = 1.74$, $p = .08$. This should call attention of the designers that PTB1 did not facilitate as reliable resolution as was achieved by PTB2.

### 3.2.6. Time-success rate and time-confidence relationships

As explained above, our purpose of analyzing the relationships between time, on the one hand, and success rate and confidence, on the other, is to examine the differences between quick and slow responses. Like resolution, time-success rate and time-confidence relationships are also calculated by within-participant correlation (or hierarchal linear regression; HLM; see Ackerman & Zalmanov, 2012). The gamma correlation between time and success rate was weaker for PTB1 ($M = .02$, $SD = .51$) than for PTB2 ($M = −.35$, $SD = .46$), $t(85) = 5.11$, $p < .0001$. In PTB1 there was in fact no correlation between them, $t < 1$ for the difference from zero, while for PTB2 it was significantly negative, $p < .0001$. For confidence, the Pearson correlation was also weaker with PTB1 ($M = −.10$, $SD = .36$) than with PTB2 ($M = −.39$, $SD = .34$), $t(94) = 6.07$, $p < .0001$, but here both correlations were significantly negative, $p < .005$ and $p < .0001$, respectively, as commonly found in the metacognitive literature.

In order to analyze the data by task sets, as done for resolution, we computed a median split of answering time for each participant in each combination of version and task set. We then compared the mean success rate and confidence for the short and long answering time (for similar analyses, see Koriat, Ackerman, Adv, Lockl, & Schneider, 2014). The means of time, success rate, and confidence per version and set are presented in Fig. 7. Differences reported above in success rates, response time, confidence, and over-confidence can be seen visually in the figure. Two interesting contributions of this presentation should be highlighted. Watching the left panel reveals that the main source for version differences in Set.1 stems from the lengthy tasks, while the differences between them in Set.2 stems from the quickly provided responses. As for the time-based slopes, it shows generally high correspondence between time-success and time-confidence slopes and reveals that the desirable state of high success rates and reliable confidence for quick and for slow responses was found only for Set.2 in PTB2. Such in-depth examination can guide future usability analysis and improvements.

Analysis of the slopes tells whether the participants benefit from investing longer in the task, or should give up as soon as they realize that the task is challenging, as they cannot improve much by investing more efforts. In order to compare the slopes, we marked by 1 cases in which lower success rate was found in the short answering time than in the longer answering time, representing a positive time-success slope. Higher success rate in the short time was marked as $-1$, representing a negative slope. Other cases were marked as 0. The same was done for time-confidence relationship. The means (and SD) of the time-success rate and time-confidence relationships and the results of the ANOVA on these values are presented in Table 2. The main effect of time—success relationship was significant, reflecting the difference between the versions reported based on gamma correlation. However, the division by task sets reveals that the relationship was negative for Set1 but not for Set2, in both versions. The difference between the versions in time-confidence relationship was marginally significant, $p = .064$ (see Table 2). The analysis by set showed a similar pattern to that of success rate, but stronger (see time-success rate and time-confidence relationship line in Table 2). The negative and positive relationships were stronger, and the case of no relationship was equivalent. It means that with Set1, in order to work efficiently participants should skip answers they cannot provide quickly, while with Set2 they can benefit from investing more effort. With this insight the designers should analyze the tasks and examine the source for this difference, aiming to allow participants to benefit from efforts with Set1 as well.

Notably, as can be seen in Table 2, with PTB1 the strength of time-confidence relationships had weaker correspondence with the time-success relationship (larger difference between the sets in the former than in the latter). This reflects lower reliability of the confidence ratings and may be the underlying reason for the weaker resolution with PTB 1 compared with PTB2.

### 3.3. Examples of usability problems uncovered by the MUP

The present study is focused on introducing the MUP methodology. Below we provide three examples of how using the MUP uniquely uncovered usability problems.

In Task 1, the participants were asked about the expected project duration. In PTB1, the main screen includes the target project duration. Many participants provided the target duration as the answer for the first task (success rate was 53%), rather than the expected project duration. They responded relatively quickly (55 s) and were quite confident in their answer (83%). As a result, overconfidence in this task was particularly high (30%). In PTB2, in contrast, no information regarding project duration is displayed on the main page. This led to longer search for the answer (73 s), but to higher success rates (73%). Confidence ratings (81%) were more in line with performance (with less overconfidence of 8%). Importantly, confidence ratings were equivalent in the two UIs, reflecting that participants were unaware of this pronounced usability difference between the UIs. This case demonstrates a situation in which one UI generates an illusion of knowing, while the other UI supports better performance and more reliable judgment.

The second example for a usability problem can be derived from Fig. 7. In the right panel, it is clear that the quick tasks were the source for the difference between the versions in Set.2. This presentation allows focusing the analysis on these quick tasks. Looking
The data for each task in each UI version revealed that the quickest task in both UIs was Task 9. In this task, the participants were asked whether the last simulation run was saved. Comparing this task performance in the two UIs revealed a pronounced response time difference between them, as well as in other measures. In PTB1, this task was performed in 43 s with 52% success rates and a small underconfidence (8%). In PTB2, it was performed very quickly (15 s) and resulted in high success rates (88%). The extent of underconfidence was comparable to the other UI (6%). On the one hand, it is encouraging that no illusion of knowing was found in this task. However, the large differences in response time and success rates should guide deeper examination of the usability of the involved features. Delving into the UIs revealed that in PTB1 opening a new scenario does not save or notify about the need to save the previous one. In PTB2, in contrast, when opening a new scenario, an appropriate notification appears and helps users avoid losing their work on the previous scenario.

The third example for the benefits that can be gained from the MUP regarding the human–computer interaction can be found in the analysis in division into quick and slow responses (Fig. 7). In particular, this analysis pointed to the tasks which the participants gave up (quick responses with low confidence), those in which they engaged in effortful but successful work (lengthy response with high success rate), and those in which they were particularly prone to overconfidence (quick or lengthy responses with low success rates). These insights can support choice of the better implementation for each task and guide the design of further improvement toward achieving higher success rates, reliable confidence, and efficient work.

4. Discussion

In the world of designing and testing software systems, a key goal of usability testing is to discover problems as well as the cognitive processes that underlie them. Understanding the source for the problems is important for guiding subsequent re-design iterations that really address the challenges users face. As we have reviewed in the Introduction, the usability testing discipline has developed various methods and tools to assess usability on the system and task levels. In order to tap task-specific usability aspects, we propose the MUP methodology. This methodology combines objective and subjective measures derived from the metacognitive research paradigm for providing a comprehensive usability picture of the compared UIs.

In the present study, we demonstrated the utilization of the MUP by comparing two UI versions of a project management tool. In particular, we demonstrated how using the MUP allows delving into the cognitive processes that takes place during user interaction with the system in a non-intrusive manner. Based on the metacognitive paradigm, we stress the importance of detailed reflection of participants’ subjective confidence. Triangulating confidence, actual success, and response time for a collection of tasks allows understanding the cognitive processing involved in working on the analyzed set of tasks with each UI.

A main message emerging from our findings is that UI should not induce illusion of success, but rather facilitate reliable confidence that supports effective regulation of efforts. Indeed, we found that the compatible set of tasks for a given UI version generated well-calibrated confidence ratings, while the incompatible task sets resulted in overconfidence. Thus, the UI of the incompatible tasks should be improved not only for achieving better performance, but also to allow participants to assess their chance for success more reliably. As for resolution, it was lower in the incompatible set in PTB1 like other measures, while PTB2 showed better resolution across both sets. This is a notable strength of PTB2, because it allowed good discrimination between the more and less successful tasks even with the less favorable task set.

Basic metacognitive research can point to potential directions for further usability testing and design improvement. For instance,
hypercorrection is the counterintuitive but robust phenomena that people succeed more in correcting errors in which they are highly confident (Butterfield & Metcalfe, 2001). Recent findings suggest that this is the case when the respondents are familiar with the correct answer, but chose to provide another answer retrieved from memory (Metcalfe & Finn, 2011). These findings imply that in advanced stages, in which the users have knowledge about using the system but nevertheless err in using it with high confidence, these errors are relatively easier to correct.

Another relevant well-established principle in the meta-cognitive literature is the “desirable difficulties” (Bjork, 1994; Bjork et al., 2013). The idea is that when learning is perceived as easy, people do not invest the effort required for maximizing their performance, while when they face a challenge, they engage in the task more and this enhances performance and improves judgment reliability. Indeed, a consistent finding with text learning tasks is that requiring in-depth processing (e.g., writing keywords summarizing the text’s essence) improves both learning outcomes and resolution (e.g., Thiede et al., 2003). Recently, calibration was also found to improve by the same strategies (Lauterman & Ackerman, 2014). In addition, knowing in advance the challenge involved in the expected test (memory of details vs. inference), allows learners to adjust their learning and judgments accordingly (Thiede, Wiley, & Griffin, 2011). Similarly, problem solvers can assess more reliably their success in an open-ended test format, which requires them to construct the solution, than in a multiple-choice test format, which allows them to use elimination of solution options (Mitchum & Kelley, 2010). These findings suggest that encouraging engagement in the task, may improve both performance and judgment reliability. Of course, in UI design, as in learning design, the principle of “desirable difficulties” should be implemented with care and after extensive testing, because while facing a challenge can be beneficial, difficulty may also hinder users’ motivation to use the system (e.g., Yue, Castel, & Bjork, 2013).

Metacognitive aspects of individual differences may be of interest for UI designers as well. It is well established that the less successful people are more prone to overconfidence than those who are more successful (e.g., Dunlosky & Rawson, 2012; Dunning et al., 2004). In the present study, no overconfidence difference was found between the novices and more knowledgeable participants and there was good calibration for the compatible set in each version (see overconfidence line in Table 2). These findings may be interpreted as suggesting on strength of the examined UIs and the selected tasks, because they did not implant illusion of knowing in the users, not even in novices, despite the challenging tasks (mean success rates of 60–85%). However, if the same tasks are used for examining experts and novices, the issue of proneness of the weaker participants to overconfidence should be taken into account, as it may harm success, efficiency, regulatory decisions, etc.

As clearly evident, the task analysis done in preparation for this study was successful in exposing differences between the two UIs, pointing to strengths and weaknesses in both. The common approach to A/B testing is obviously to have participants perform typical tasks. By having our selected task set divided into two subsets of tasks, each more appropriate for one of the UIs, we increased the sensitivity of the test and found differences that would not be uncovered otherwise. We recognize that such an approach worked well in this study. However, one may wonder how such an approach can generalize to other usability tests. Specifically, should there be such specific task analysis in any usability testing? The choice of test tasks is critical in usability evaluation, as was shown before by several studies (Cockton & Woolrych, 2001; Hertzum & Jacobsen, 2003; Lindgaard & Chattratichart, 2007). Our findings also highlight task choice as a crucial step that affects the findings and directions for improvement. Although using the most typical tasks does fit the MUP methodology as well, our study suggests that usability researchers should consider also differentiating tasks by their suitability to different UIs for exposing important usability strengths and weaknesses.

However, we waived two considerations that testers implementing the MUP should take into account. First, we used only 4–5 tasks in each set. The reason for this low number of tasks was practical, rather than theoretical. We wanted to have a within-participant comparison between the two UIs and thus a session took about two hours. We demonstrated how insights regarding relationships among the various measures can be obtained even with this number of tasks (see above resolution, time-success rate, and time-confidence relationships). However, using more tasks would allow straightforward within-participant by correlations and regressions, which is more in line with the metacognitive literature (e.g., Ackerman & Zalmanov, 2012; Metcalfe & Finn, 2008) and would also increase the statistical power of other measures. This can be achieved by a between-participant design or by inviting participants for more than one meeting. Second, we did not make the two sets comparable in any aspect. This precluded comparing the sets in the various ANOVAs. By this choice we demonstrate degrees of freedom in the design of the usability test. However, cases that allow making comparable sets may expose differences between the interfaces that we could not detect.

Another crucial step is the selection of target population. We chose novice and more knowledgeable undergraduate students because the tested systems accompany academic courses. However, these systems allow also real-life project management. For examining their commercial value, usability testing by experienced project managers is clearly called for. This requires redoing the task analysis with a sample from the target population, and changing the task sets used for the comparison accordingly. The tasks should not be too easy or too hard for the target population, because this will generate ceiling or floor effects, respectively, either objectively (success rates) or subjectively (confidence). Both types of variability are required for finding differences and thus points for improvement.

In summary, the system-level measures regarding the full project management task (task 10) and the SUS questionnaires showed similar usability between the two compared UIs. In this study, we suggested additional measures that can be collected easily, without disrupting users in their interaction with the system, for drawing more fine-tuned usability conclusions. Based on insights from the metacognitive approach, we propose the MUP as a comprehensive methodology for comparison among similar systems. As demonstrated in this study, the MUP is highly sensitive to the possibly that some functions are more effective and some are less effective within a given UI. Thus, the detailed analysis allows allocating in each UI the functions that work better. The deep understanding of strengths and weakness in the existing UIs is a powerful basis for designing the next version of the system and for developing other systems. Moreover, understanding the challenges involved in using a chosen UI may guide appropriate focus in user training programs and in guidelines for support personnel.

Appendix. Tasks
The list of the nine tasks used to examine the compared user interfaces in the order they were employed. The table lists for each task the preferable version and the usability issue in the less preferable version.

<table>
<thead>
<tr>
<th>Task</th>
<th>Preferable version</th>
<th>Usability issue in the less preferable version</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. What is the estimated duration of the project?</td>
<td>PTB.2</td>
<td>Misleading duration display</td>
</tr>
<tr>
<td>2. Does the estimated duration meet the end time requirement?</td>
<td>PTB.2</td>
<td>Misleading duration display</td>
</tr>
<tr>
<td>3. For each of the following five tasks, list the preceding tasks. (five project task identifiers were listed)</td>
<td>PTB.2</td>
<td>Scroll is required for examining the entire network</td>
</tr>
<tr>
<td>4. How much money will you have at the end of week 23 of the project?</td>
<td>PTB.1</td>
<td>The information is absent from the main screen and the design of the icon leading to this information is confusing</td>
</tr>
<tr>
<td>5. How many employees of each type (e.g., engineers, mechanics) are required for starting the project?</td>
<td>PTB.1</td>
<td>The information regarding employees and their division between tasks is hidden</td>
</tr>
<tr>
<td><strong>Type</strong></td>
<td><strong>Number</strong></td>
<td></td>
</tr>
<tr>
<td>6. How many employees are required for each of the tasks to be performed at the initial phase of the project?</td>
<td>PTB.1</td>
<td>Same as in task 5</td>
</tr>
<tr>
<td>Task</td>
<td>Employee Type 1</td>
<td>Type 2</td>
</tr>
<tr>
<td>7. Hire employees as you suggested in Task 5</td>
<td>PTB.1</td>
<td>The information is not visible in the main screen. In addition, the icon leading to this option and the pop-up window are not clear</td>
</tr>
<tr>
<td>8. Open a new scenario named LRMB located on the Desktop</td>
<td>PTB.1</td>
<td>The icon is hidden behind the software logo and its design is not clear</td>
</tr>
<tr>
<td>9. Was the last simulation run saved?</td>
<td>PTB.2</td>
<td>While opening a new scenario there is no warning message that the current run will be lost</td>
</tr>
</tbody>
</table>

References


