Mental Workload at Transitions between Multiple Devices in Personal Information Management

Manas Tungare
Virginia Tech / Google
manas@tungare.name

Manuel A. Pérez-Quiñones
Virginia Tech
perez@cs.vt.edu

ABSTRACT
Knowledge workers increasingly use multiple devices such as desktop computers, laptops, cell phones, and PDAs for personal information management (PIM) tasks. This paper presents the results of a study that examined users' mental workload in this context, especially when transitioning tasks from one device to another. In a preliminary survey of 220 knowledge workers, users reported high frustration with current devices' support for task migration, e.g. accessing files from multiple machines. To investigate further, we conducted a controlled experiment with 18 participants. While they performed PIM tasks, we measured their mental workload using subjective measures and physiological measures. Some systems provide support for transitioning users' work between devices, or for using multiple devices together; we explored the impact of such support on mental workload and task performance. Participants performed three tasks (Files, Calendar, Contacts) with two treatment conditions each (lower and higher support for migrating tasks between devices.)

Workload measures obtained using the subjective NASA TLX scale were able to discriminate between tasks, but not between the two conditions in each task. Task-Evoked Pupillary Response, a continuous measure, was sensitive to changes within each task. For the Files task, a significant increase in workload was noted in the steps before and after task migration. Participants entered events faster into paper calendars than into an electronic calendar, though there was no observable difference in workload. For the Contacts task, time-on-task was equal, but mental workload was higher when no synchronization support was available between their cell phone and their laptop. Little to no correlation was observed between task performance and both workload measures, except in isolated instances. This suggests that neither task performance metrics nor workload assessments alone offer a complete picture of device usability in multi-device personal information ecosystems. Traditional usability metrics that focus on efficiency and effectiveness are necessary, but not sufficient, to evaluate such designs. Given participants' varying subjective perceptions of these systems and differences in task-evoked pupillary response, aspects of hot cognition such as emotion, pleasure, and likability show promise as important parameters in the evaluation of PIM systems.

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User interfaces—*Evaluation/ methodology*

Author Keywords
Personal Information Management, Mental Workload, Multiple Devices

INTRODUCTION

As we amass vast quantities of personal information, managing it has become an increasingly complex endeavor. The emergence of multiple information devices and services such as desktops, laptops, cell phones, PDAs and cloud computing adds a level of complexity beyond simply the use of a single computer. It is common for a lot of people to carry a laptop computer or a cell phone as they go about their everyday business, outside the usual contexts of an office or a home [7, 35], and to expect productive work output when mobile. However, the current state-of-the-art in information management solutions sends these users into a frenzy trying to locate the most current version of their slide shows, the documents they sent around for review, and the phone number of the person they need to call right now. In traditional single terminal computer systems, the majority of a user's attentional and cognitive resources are focused on the terminal while performing a specific task. However, in an environment where multiple devices require intermittent attention and present useful information at unexpected times, the user is subjected to different mental workload.

Earlier, we conducted a survey study [35] to understand the use of multiple devices in personal information and identify common tasks, activities, devices, patterns, device affinities, and problems in their use. Many findings were as expected: that users preferred laptop computers over desktops; several users owned and regularly used more than two computers, plus a cell phone, a digital camera, etc. However, a surprisingly high number of users reported chronic problems in using multiple devices together for managing their tasks. Synchronization issues between information collections on two or more machines were cited as the most common problem. Sprouting from this investigation, we decided to examine this problem deeper—whether the level of system support for such basic processes as information migration affects user performance and workload.

In the survey, several users were very passionate in reporting horror stories of their use of multiple devices. Many of them had faced issues ranging from not being able to contact a person when they needed to, to complete data loss when transferring data between devices. The tone of the narration of their experiences in response to the questionnaire revealed a

deep undercurrent of frustration at the status quo in personal information management tools. While current usability metrics are able to provide evaluations of interfaces based on objective qualities such as efficiency and performance, other non-traditional factors such as user enjoyment, acceptance, happiness and satisfaction are neither measured nor reported in studies.

MOTIVATION

Content analysis of the survey responses revealed that many of the issues that users faced could be studied and understood within the framework of mental workload. E.g. factors such as frustration level, mental demand and perceived ratings of own performance are all dimensions of the NASA TLX scale. It has been shown that an operator's task performance is inversely correlated with high levels of mental workload [24]. Thus, we set out to explore if mental workload estimates could be used to compare task difficulty in PIM tasks. Prior work in mental workload measurement has established that physiological measures such as changes in pupillary diameter (known as Task-Evoked Pupillary Response [3]) can be used to estimate mental workload. Such continuous measures of mental workload can help locate sub-tasks of high task difficulty. Iqbal et al. [15] demonstrated that within a single task, mental workload decreases at sub-task boundaries. A fundamental goal of our research was to examine if their finding still applies when the second sub-task is performed on a different device than the first. Our contrary hypothesis was that mental workload rises just before the moment of transition, and returns to its normal level a short duration after the transition is complete.

The specific research questions were as follows:

RQ1. Mental Workload and Support for Multiple Devices

What is the impact of (1) different tasks and (2) different levels of system support for migrating information, on the workload imposed on a user? Certain tasks require more attentional resources than others, and may result in increased mental workload, while certain other tasks may be straightforward and may require fewer mental resources. What is the variability in the subjective assessment of mental workload for these tasks?

Systems differ in the level of support they provide for pausing a task on one device, and resuming it on another [31]. A goal of our research was to examine if mental workload at the point of transition was correlated with the level of system support available for the sub-task of transitioning. Miyata and Norman hypothesized [22] and Iqbal et al. [15] demonstrated that within a single task, mental workload decreases at sub-task boundaries. But when a sub-task is performed on a different device than the first, what are the changes in mental workload?

RQ2. Operator Performance & Levels of System Support

How is user performance impacted at differing levels of system support for performing tasks across multiple devices? To evaluate this, we simulated two conditions for each task; in each case, the L0 condition offered a lower level of support for migrating tasks between devices than the L1 condition. How does operator performance in condition L0 compare to that in condition L1? Several measures of task performance were used, on a per-task basis. Many of these are commonly used in traditional usability evaluations as well, e.g., mean time on task, number of errors.

RQ3. Operator Performance and Mental Workload

Are subjective assessments of mental workload an accurate indicator of operator performance in this domain? Are both, subjective measures of workload (NASA TLX) and the physiological measure (pupil radius), sensitive to workload in PIM tasks? It is clear that workload does not stay constant during a task, but varies constantly. What are the types of changes that can be observed in workload during the execution of a task? How do the two measures of workload each correlate with task performance? Mental workload has been shown to be negatively correlated with several of these metrics in other domains [24, 1, 4]. Does the same (or a similar) relationship hold between mental workload and task performance in the PIM domain?

RELATED PRIOR WORK

Personal Information Management

This work overlaps three broad areas: Personal Information Management (PIM), Multi-Device Interfaces and Mental Workload Measurement. Studies in PIM include investigations of individual collections such as files [2], calendars [18, 27, 34], contacts [38], email [39, 12], bookmarks, etc. as well as users' information management practices [21], using a range of investigation techniques [32]. Issues such as information overload and information fragmentation [5] have also received attention. However, the issue of information fragmentation across multiple devices [17] looms larger as mainstream users increasingly have started to use portable devices such as cell phones, portable digital assistants (PDAs) and laptop computers for PIM.

PIM using Multiple Devices

In prior work [30], we explored the issues that arise in multi-device interfaces, especially when several devices are used together to perform a single task. The flow of information among a user's multiple devices has been likened to a biological ecosystem [28]. Several concepts in Personal Information Ecosystems are analogues of related concepts from biological ecosystems, and the metaphor helps construct a meaningful information flow among devices. While task migration is handled at the interface level, seamless data migration requires system support. The Syncables framework [36, 37] was developed in response to the need for being able to access data from any of a user's devices without extraneous task steps. It has been recognized widely that the mobile context is fundamentally different from the stationary context [25], and design must therefore account for the differences [29]. Dourish [9] refers to situated interaction as *"embodied interaction"*, and outlines several principles that designers must take into account for technology that, by its very nature, must co-exist in the environment that users use it in.

Holistic Usability in Multi-Device Environments

The origins of usability and human factors can be traced back to factories and environments where users performed specific duties at specific times. The goal of human factors specialists was to optimize operator performance and the fit between human and machine. Modern developments in the science of cognition have examined the relationship of the user in complex computing environments, and place greater emphases on the situational aspects of human-computer interactions. Distributed cognition theory [14] extends the reach of what is considered cognitive beyond the individual to encompass interactions between people and with resources and materials in the environment. In multi-device computing environments, it is worthwhile to analyze the system as an integrated whole whose purpose is to assist the user in satisfying her information needs. Other recent theories such as Embodied Interaction [9] also support the notion that technology and practice are closely intertwined; they co-exist and co-evolve.

Hot Cognition Aspects in the Evaluation of PIM

Norman [23] argues that emotion plays a central role in our interaction and appreciation of the computing devices we use. But classic usability metrics fail to account for subjective factors such as emotional appeal, frustration, and likability. All these point to the necessity of bringing hot cognition aspects into the evaluation process: Jordan [16] advocates designing for pleasurability of the user, stating a hierarchy of needs for a computing system: functionality as the most basic, then usability, and finally, pleasure. Thus, usability is necessary but not sufficient to guarantee an optimal user experience. Kelly et al. [19] identify a shortcoming in PIM studies as well; quality of life measures (e.g. [11]) have received received little attention in PIM evaluations.

Mental Workload Assessment

Mental workload is defined as *"that portion of operator information processing capacity or resources that is actually required to meet system demands"* [24, 10]. It is task-specific and operator-specific (i.e., person-specific); the same task may evoke different levels of workload in different individuals. *Task complexity* is related to the demands placed on an operator by a task, and is considered operator-independent, whereas *task difficulty* is an operator-dependent measure of perceived effort and resources required to attain task goals [6]. Mental workload is considered an important, practically relevant, and measurable entity [13]. Several ways of measuring mental workload are used in practice: Performance-based Assessment Techniques; Subjective Workload Assessment Techniques, e.g. NASA Task Load Index (TLX) [13]; and Physiological Workload Assessment Techniques, e.g. task-evoked pupillary response [3, 20].

METHODOLOGY

Representative Tasks

From a content analysis of survey data, the following emerged as the most common tasks:

- **File Synchronization.** One of the most commonly reported frustrating tasks that emerged was synchronizing data (this echoes findings by others [7]). Users' responses to this question elicited a long list of problems and issues that they often encountered.

  Participants were asked to play the role of a consultant who worked with several clients, either at their own office on the desktop computer, or at one of the clients' sites, using their laptop. On each machine, an exact replica of a file system was provided, either deeply-nested, moderately-nested, or flat, based on participant preferences. Instructions were provided, one at a time, asking them to make certain specific edits to files. Mid-way, they were asked to wrap up their work and travel to a client site. In L0, they were provided USB drives and web-based email; in L1, a network drive allowed remote access to files.

- **Managing Calendars.** One of users' main motivations for using more than one device was to be able to access their calendar information when away from their desks. The use of paper calendars is widespread, even despite the availability of online calendars. It is not clear which of these methods is easier; almost equal numbers of participants reported preferring one over the other for several reasons [34].

  At the start of the calendar task, users were provided either two paper calendars labeled 'Home' and 'Work' (L0) or an online calendar program with two overlapping calendars in it, also labeled 'Home' and 'Work' (L1). During the task, participants were presented instructions that required them to consult or update their calendars. Different types of events included tentative, rescheduled, group events, events that required preparation, and conflicting events (details in [33]).

- **Contact Management.** Contact management on phones was identified as a frustrating task due to deficiencies in the phone interface (n=5), or a lack of features in the specific software they used, both on the computer as well as on the phone (n=3).

  Participants were described a scenario where they were a researcher attending a conference, and met several old acquaintances and made new contacts. They were allowed to access their laptop at some times, and their phone at other times, and both at some other times.

Experiment Design

In this experiment, we were interested in the impact of two factors—task, and level of support—on workload in participants. Since individual differences in work practices, task performance, and assessments of workload would display high variability across participants, a within-subjects design was used. Each participant was assigned to each cell, making this a complete block design (at 3×2 treatment levels). Each experimental task identified above was assigned to users to be performed in one of two sessions separated by at least two weeks, in order to minimize the learning effects associated with the first session. The order of tasks was completely counterbalanced. Figure 1 shows a graphical overview of the entire experimental setup.

Pilot studies were conducted with five participants. Training

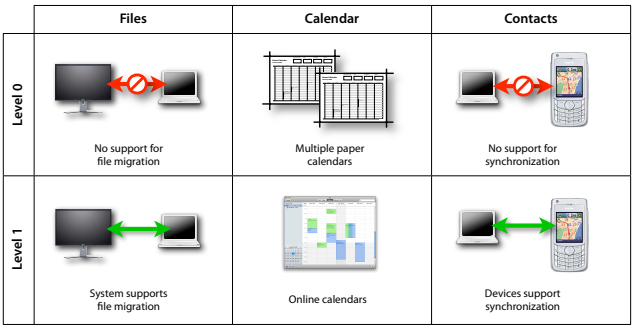

Figure 1. An overview of experimental tasks

was provided in the form of demonstration videos, hands-on time, and required completion of a set of 10 familiarization tasks. Sample size estimation conducted after 6 participants had performed the experiment revealed that a medium to large effect was evident according to Cohen's *d*. The sample size chosen was 21, higher than that required to detect such an effect with a power of 0.8 at the α=0.05 level of significance for all three tasks, and to allow for experimental mortality (since it was conducted in two sessions.)

Participants were presented with a desktop computer, a laptop and a cell phone. Between the two computers, instructions were presented on a large 30-inch display. A custom web application was written to present instructions to the participants, one at a time. When the display changed from one instruction to the next, the app recorded the timestamp. This was later used to analyze sub-task-level changes in physiological measures of mental workload. Participants were requested to provide a subjective estimate of workload using the NASA TLX scale after each task. Pupil radius measurement was performed using a mobile head-mounted eye tracker. Illumination was carefully controlled to be the same for all participants and at all times of the day. The experiment was conducted in a closed room, and no external light was allowed to enter the room. The raw pupil data was extremely noisy and needed to be smoothened to isolate the signal from the noise, using the Savitzky-Golay filter. After smoothing, pupil radius data was adjusted to account for individual differences in pupil size. A baseline reading for pupil radius was obtained for each participant from the first 5 seconds of pupil activity data. During the first five seconds, participants were not assigned any specific task or provided any instructions to read, and was considered a period of minimal task-induced workload.

RESULTS

Results for Research Question 1

Research Question 1 explores the impact of (1) different tasks and (2) different levels of system support for migrating information, on the workload imposed on a user.

Subjective Metrics using NASA TLX

From an ANOVA of NASA TLX scores, Task was seen to have a main effect on Overall Workload (OW) ($F_{(2,102)}$=4.75; p=0.011). Post hoc analysis using Tukey's HSD showed that the Contacts task imposed significantly lower overall workload than the Files task (p=0.0074). Level of support for performing tasks across multiple devices (L0 vs L1) did not influence Overall Workload and there were no significant interactions.

This suggests that while NASA TLX ratings are able to discriminate between different tasks in the personal information management domain, the scale is not sensitive enough to detect differences in performing a task using two or more techniques. One reason for this could be that NASA TLX, being a subjective measure, can only be administered at the end of a task. It thus fails to capture variation in workload within a task, and provides only an aggregate per-task measure of workload.

Mean (SD)	Files	Calendar	Contacts
L0	41.11 (20.85)	36 (18.80)	30.89 (16.65)
L1	38.61 (18.92)	31.17 (18.91)	22.89 (11.49)

Table 1. Means (SDs) of Overall Workload ratings

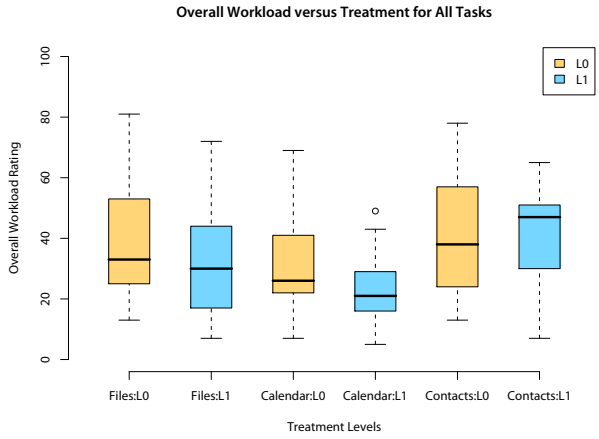

Figure 2. Overall Workload across Treatments

Similar effects were seen for three individual dimensions of the NASA TLX scale as well:

- **Mental Demand.** Task had a main effect on Mental Demand (MD) ($F_{(2,102)}$=6.69; p=0.0019). Post hoc analysis results for Mental Demand using Tukey's HSD revealed that the Files task imposed significantly higher Mental Demand than the Contacts task (p=0.0024), similar to the effect seen in case of Overall Workload.
- **Frustration.** Task had a main effect on subjective reports of frustration provided by participants ($F_{(2,102)}$=6.57; p=0.0021). Participants noted significantly higher frustration ratings for the Files task as compared to the Contacts task (p=0.0014, using Tukey's HSD). Differences among the other two pairs (Files-Calendar and Calendar-Contacts) were not significant.
- **Own (Perceived) Performance.** In this dimension, lower numbers indicate better performance. Participants rated

their Own Performance differently for the three task conditions ($F_{(2,102)}$=3.37; p=0.038).

Task-Evoked Pupillary Response

For the Contacts task, significant differences were found for each step between the two levels of system support in task migration (synced versus unsynced conditions.) Graph 3 illustrates the means (SDs) and p-values for each step.

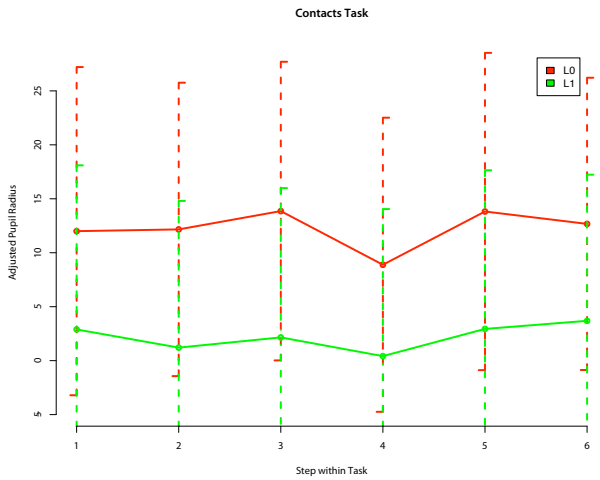

Figure 3. Adjusted pupil radius for each step of the Contacts task.

Differences in TEPR Between Steps in the Same Task

In the Files task, Level 0 (where participants used USB drives or email-to-self), significant differences were noted in the workload for the steps before and after the migration step ($F_{(8,136)}$=7.8835; p=1.12×10^{-8} using Tukey's HSD). This suggests that there is a distinct increase in workload before and after the migration step, when there is a lack of support for task migration. It is interesting to note that no significant differences were found in the L1 condition for the same task, suggesting that the file migration support has an effect on differences in workload before/after migration.

TEPR within Critical Sub-Tasks

Graphs 4 & 5 depict the task-evoked pupillary response for several participants for the Files task. These are time-series graphs (time in seconds on the X axis) against adjusted percent pupil radius on the Y axis. In the Files task, Step 5 was the critical task migration step, in which participants were required to pause their task on the desktop and to move to the laptop. As can be seen, the task-evoked pupillary response (TEPR) rises soon after the start of the critical step, and reaches a (local) maxima. In some instances, it progressively lowers, and in some, it stays at the new, higher level of workload until the end of the task. This provides support for the hypothesis that steps that involve transitions between devices lead to high mental workload.

Summary of RQ 1 Results

In NASA TLX scores, Task was seen to exhibit a main effect on Overall Workload, Mental Demand, Frustration and

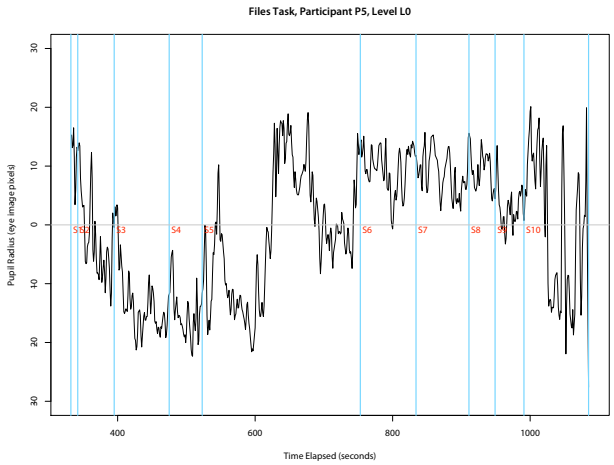

Figure 4. Task-evoked pupillary response, Participant P5, Files Task, L0

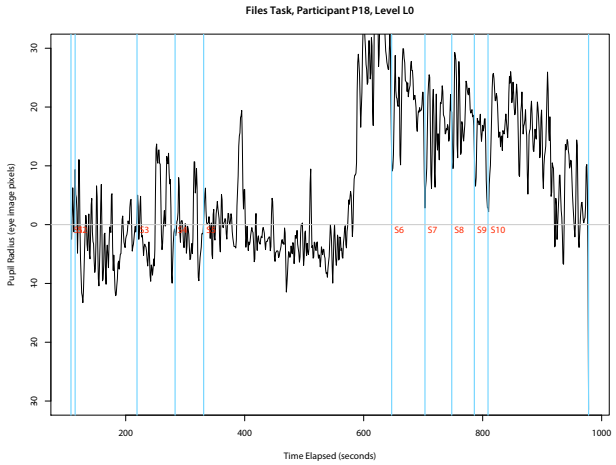

Figure 5. Task-evoked pupillary response, Participant P18, Files Task, L0

Own Performance, but not on the other three scales. There was no difference seen on any scale between two treatment levels of the same task. This suggests that NASA TLX is not very sensitive to changes in workload in the kinds of personal information management tasks tested in this experiment. Because of its lack of ability to discriminate between two or more ways of performing the same task, its validity and usefulness in PIM tasks cannot be established with the evidence obtained.

Task-evoked pupillary response, on the other hand, provided important insights into task migration. Specifically, it showed a significant difference for each step of the Contacts task between levels L0 and L1. Also, it showed significant differences between pre- and post-task-migration steps in the Files task. It was observed from the data that local maximas were attained during the task migration step. All of this points to the potential usefulness of task-evoked pupillary response as a continuous measure of workload in PIM tasks.

Results for Research Question 2

Research Question 2 seeks to explore the differences in operator performance, if any, between the L0 and L1 task conditions. The primary measure of operator performance used in this study (for all tasks) was time on task. Others, such as number of errors, number of entries made, etc. were defined, measured and evaluated on a per-task basis. For the Files and Calendar tasks, no significant differences were found in the time taken to complete the task. However, for the Contacts task, participants completed the task significantly faster in the presence of synchronization support than without ($F_{(1,34)}$=4.72; p=0.037).

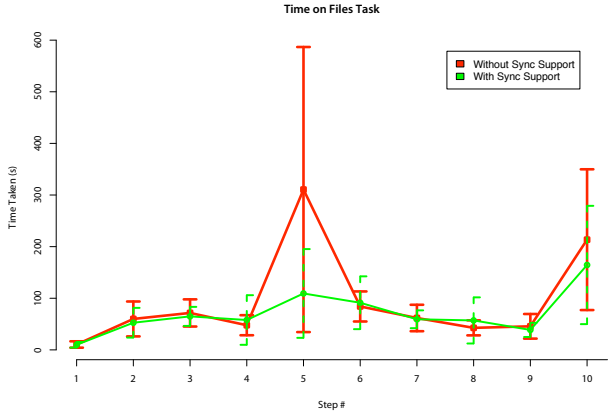

Figure 6. Time on task, per Step, in the Files task.

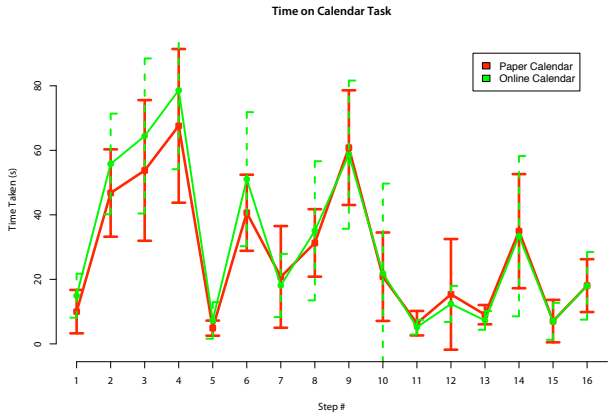

Figure 7. Time on task, per Step, in the Calendar task.

Significant differences ($F_{(1,34)}$=8.83; p=0.0054) were found for the transitional step in the Files task (Step 5) where participants were requested to pause work on their desktop computers and resume it on a laptop, taking their files with them, but not for any other step. This was expected; in fact, the lack of significant differences for steps that did not involve a transition from one device to another in the Files task confirms that the experimental setup did not lead to any biases in steps that were identical by design in both treatment levels.

For the Calendar task, two steps took significantly different

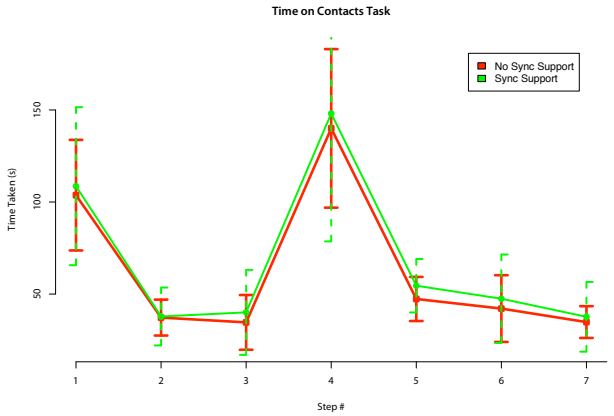

Figure 8. Time on task, per Step, in the Contacts task.

times in case of the paper calendars versus online calendar ($F_{(1,34)}$=4.33; p=0.045). Both steps involved proposing a meeting time and scheduling it on the calendar. In both instances, participants took lesser time using a paper calendar than an online calendar. The ease of quick capture in paper calendars might explain why it is the tool of choice for several users despite the widespread availability of online calendars.

Participants correctly edited more files ($F_{(1,34)}$=5.52; p=0.025) in the condition with no support for file synchronization (Mean=6.40; SD=0.92 files) than in the condition with synchronization (Mean=5.22; SD=1.90 files) from a maximum of 7 files. This was an unexpected finding, disproving Hypothesis 2 (at least for one particular task metric) that task performance would be higher in the L1 condition.

In contact management, the number of entries made on the secondary device was significantly different in both treatment levels ($F_{(1,32)}$=15.86; p=0.00037): participants who managed contact information with syncing support made 4.71 entries on the other device, while participants without such support made only 1.00 entries. (If an instruction clearly required participants to add a contact record to a specific device (either the laptop or the phone), that device was termed the primary device. The other device (either the phone or the laptop, respectively) was termed the secondary device.)

Summary of RQ 2 Results

For the Files task, the time taken to perform the critical step in the Files task — moving from the desktop to the laptop — was significantly higher when there was a lack of system support for such migration (implemented in this experiment as a Network Drive). However, more files were edited correctly in the case where synchronization had to be performed using USB drives or email-to-self. For Calendars, there was no difference in any task metrics between the paper and online calendar conditions. In the Contacts task, more entries were recorded on secondary devices when synchronization was automatic. Thus, little to no support was found for Hypothesis 2, especially with the observation that more files

were edited correctly with lower levels of support for task migration.

Results for Research Question 3

Research Question 3 examines if measures of mental workload may be used as predictors of task performance in personal information management tasks. Since time-on-task was the only performance metric that was (1) used for all three tasks, and was (2) not subject to any ceiling effects, further analysis of the correlation between performance and workload focuses on this metric. Mental workload was estimated via two methods; we consider them separately to examine whether either or both of them may be used as task performance predictors.

NASA TLX Ratings as Predictors of Operator Performance

Significant correlations were seen between NASA TLX subscales and time-on-task only in the following isolated cases: Overall Workload for Files Level L1 (p=0.01, r=0.57), Mental Demand for Files Level L1 (p=0.0071, r=0.61), Own (Perceived) Performance for Files L0 (p=0.05, r=0.47), Own (Perceived) Performance for Files L1 (p=0.02, r=0.54), Frustration for Files L0 (p=0.05, r=0.47), Frustration for Calendar L0 (p=0.51, r=0.17).

Task-Evoked Pupillary Response as a Predictor of Operator Performance

Workload estimated according to the Task-Evoked Pupillary Response was not found to be significantly correlated with Time on Task, using Pearson's product-moment coefficient (r). Table 2 shows the correlation coefficients and p-values for each task condition. It can be inferred that mental workload (measured via pupillary response) is not a good predictor of task performance.

TEPR × Time	L0	L1
Files	r=-0.062, p=0.46	r=0.15, p=0.063
Calendar	r=-0.11, p=0.078	r=-0.067, p=0.283
Contacts	r=-0.13, p=0.18	r=0.042, p=0.68

Table 2. Pearson's *r* for Task-Evoked Pupillary Response for each task condition.

Summary of RQ 3 Results

Neither NASA TLX ratings nor task-evoked pupillary response showed consistent correlation with task performance. Isolated instances of significant correlations were observed, but they do not support the use of workload measures as predictors of task performance. The lack of any meaningful correlation between performance-based metrics and workload metrics suggests that neither alone is sufficient to assess and describe highly contextualized tasks in the domain of personal information management. Thus, Hypothesis 3 was disproved in case of both metrics used in the measurement of mental workload.

Other Observations

While the preceding sections provide answers to the research questions posed at the start of this study, there were several interesting observations noted while participants performed the experimental tasks.

- **Lack of Preparation in Task Migration.** None of the participants performed any kind of planning tasks at the start of the Files task to prepare for migration. Since the means of task migration (USB drives, email access and network drive access) were already provided to them, it would have been possible for them to plan ahead by copying their files to the network, for example. However, none did so.

  This lack of planning has significant implications for those designing technologies for mobility: users cannot be expected to plan ahead or to prepare for a device transition [29]. Task migration technologies must take into account the opportunistic use of multiple devices without any pre-planning and must initiate any pre-migration activities without the need for explicit user intervention [30].

- **Maintaining Contextual Awareness in Calendars.** In the Calendar task, a few of the instructions provided to the participants mentioned the current date as a way to anchor them in temporal context. Since an entire week's worth of calendar events were presented in about 10 to 15 minutes, it was important to introduce the current day in order to preserve the hypothetical temporal unfolding of events in the experimental tasks. Participants adopted various techniques to maintain this temporal context while interacting with the calendars. Those who used the electronic calendar clicked the specified date in the calendar window, which would then highlight that day in the display. Such a visual representation helped as an external cognition aid so that the task of remembering the current day could be offloaded to the environment. Very few users who used paper calendars used similar techniques: those that did, marked each passing day with a dot or a cross towards the top of the day.

- **Capturing Information about Tentative Events in Calendars.** The scheduling of tentative collaborative events caused a high amount of confusion to users (noted via experimenter's observations; not statistically significant). Using multiple paper calendars, participants indicated the changes and rescheduling with an assortment of arrows, scratched lines, and other idiosyncratic annotation techniques. In electronic calendars, while participants could reschedule an event easily by dragging-and-dropping the electronic representation of the event to the rescheduled time, this did not solve the entire problem.

  The larger issue in tentative collaborative events is the ad hoc specification of attendees' constraints. Current calendar systems do not capture the set of constraints that lead to the tentative scheduling of an event. Hence, when such an event is to be moved to another time, the new start time must be evaluated against the complete set of constraints by consulting the originating source, e.g. email. The event record within an electronic calendar provides no way to indicate the justification behind the particular choice of time, and thus lacks an affordance for potential rescheduling. This is also a problem when adding a new constraint to the mix.

While a few calendar systems do provide support for automatic multi-party meeting scheduling, the resulting artifact is a calendar event, not an expression of the constraints. This makes it difficult to add or remove constraints from the mix, to arrive upon a different time than originally scheduled.

DISCUSSION

Through the results of these studies, I found that specifics of the tasks and levels of support for task migration affected users' perceived workload ratings as well as task-evoked pupillary response in a variety of ways. These workload metrics were not the traditional usability metrics that are often used to evaluate computing systems, such as performance, efficiency, errors, etc. In fact, metrics such as whether users were able to answer questions correctly and time-on-task showed little to no difference with the different ways of performing a task, with and without support for task migration.

What this points to is that while both types of systems result in similar outcomes (and thus would be rated equally on traditional usability metrics), they do not evoke the same experiences in users. Frustration, mental demand, and workload: all are components of the entire user experience, but are not often captured by researchers and designers when assessing personal information ecosystems. This points to two separate, yet related, issues that warrant discussion: (1) evaluating usability using concepts from hot cognition that are more representative of user concerns when using multiple devices together, and (2) evaluating usability for a device ecosystem together instead of as disparate devices.

Evaluating Usability using Hot Cognition Aspects

Besides the need to measure traditional usability metrics, it is important to test whether we are, in fact, measuring the right metrics. Dillon notes [8] that in several tasks, efficiency may not be the user's priority. In particular, he highlights the inadequacy of traditional usability measures for many high-level, ongoing tasks such as information retrieval and data analysis. Other studies also have shown [26] that users' preferences for particular brands of devices have significant effects on their perception of usability of those as well as other devices. This shows that aspects of hot cognition such as affect, emotion, personal preferences, etc. play an important role in the user experience — perhaps an even greater role than purely objective metrics such as task completion times and feature comparisons.

Holistic Usability for Personal Information Ecosystems

Distributed cognition theory recognizes that actors in a system often rely on the use of external artifacts to augment their own cognition. Usability cannot thus be embedded into an artifact, but is distributed across an entire activity system. This is evident in this study in various ways: users performing the Calendar task kept track of the current day by highlighting that day in an online calendar, or by marking off corresponding days in a paper calendar. In the Files task, a few users kept modified files open in their respective editor programs as a means of tracking their changes. While these are just a few idiosyncratic examples, it points to the larger issue of systems and devices lacking explicitly-designed support for external cognitive tasks.

CONCLUSIONS

Pure performance-based measures are not sufficient to describe and assess highly contextual tasks in the domain of personal information management, and the inclusion of user perception in their assessment is important. Traditional usability metrics emphasize efficiency, effectiveness and satisfaction [ISO 9241], but they relegate metrics such as pleasure and emotion to the sidelines. This study describes that while performance metrics do not show much difference, mental workload (measured via the task-evoked pupillary response) shows a difference with/without support for synchronization (in the Contacts task).

Many devices that are intended to be used in collaboration with other devices are designed independently of one another. In some cases, it appears as if minimal attention has been given during the design process to understand the broader context of use and to situate the device in this context, offering support for the activities that are performed in real use scenarios. When evaluated for usability, many devices are often tested in pristine laboratory settings. Even if tested in real world scenarios, they may not be evaluated together with other interacting devices in the user's work environment. The lack of correlation in this experiment between task metrics and workload measures stresses the need for conducting holistic usability evaluations of such devices when they act together to fulfill a user's information needs.

ACKNOWLEDGMENTS

We would like to thank Tonya L. Smith-Jackson for instigating some of the ideas behind this project. Steve Harrison, Edward A. Fox, Stephen Edwards, Pardha S. Pyla, and Ranjana Mehta, all provided important insights that led to the design of this study in its current form. Thanks are also due to our participants who spent considerable time in completing both sessions of the experiment.

REFERENCES

1. J. Ballas, C. Heitmeyer, and M. Pérez-Quiñones. Evaluating two aspects of direct manipulation in advanced cockpits. In *CHI '92: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 127–134, New York, NY, USA, 1992. ACM.

2. D. Barreau. Context as a factor in personal information management systems. *Journal of the American Society for Information Science*, 46(5):327–339, 1995.

3. J. Beatty. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, 91(2):276–92, 1982.

4. D. A. Bertram, D. A. Opila, J. L. Brown, S. J. Gallagher, R. W. Schifeling, I. S. Snow, and C. O. Hershey. Measuring physician mental workload: Reliability and validity assessment of a brief instrument. *Medical Care*, 30(2):95–104, 1992.

5. R. Boardman, R. Spence, and M. A. Sasse. Too many hierarchies?: The daily struggle for control of the workspace. In *Proc. HCI International 2003*, 2003.

6. D. de Waard. *The Measurement of Drivers' Mental Workload*. PhD thesis, University of Groningen, Haren, The Netherlands, 1996.

7. D. Dearman and J. S. Pierce. "It's on my other computer!": Computing with multiple devices. In *CHI 2008: Proceedings of the ACM Conference on Human Factors in Computing Systems*, page 767, 2008.

8. A. Dillon. Beyond usability: process, outcome and affect in human-computer interactions. *Canadian Journal of Library and Information Science*, 26(4):57–69, 2002.

9. P. Dourish. *Where The Action Is: The Foundations of Embodied Interaction*. MIT Press, October 2001.

10. F. T. Eggemeier, G. F. Wilson, A. F. Kramer, and D. L. Damos. *Workload assessment in multi-task environments*, chapter 9. Multiple-Task Performance. Taylor and Francis, 1991.

11. J. Endicott, J. Nee, W. Harrison, and R. Blumenthal. Quality of life enjoyment and satisfaction questionnaire: A new measure. *Psychopharmacology Bulletin*, 29(2):321, 1993.

12. J. Gwizdka. Email task management styles: The cleaners and the keepers. In *CHI '04: Extended Abstracts on Human Factors in Computing Systems*, pages 1235–1238, New York, NY, USA, 2004. ACM Press.

13. S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human Mental Workload*, 1:139–183, 1988.

14. E. Hutchins. *Cognition in the Wild*. MIT Press, 1995.

15. S. T. Iqbal and B. P. Bailey. Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption. In *CHI '05: Extended Abstracts on Human Factors in Computing Systems*, pages 1489–1492, New York, NY, USA, 2005. ACM.

16. P. W. Jordan. *Designing Pleasurable Products: An Introduction to the New Human Factors*. CRC, 2000.

17. D. Karger and W. Jones. Data unification in personal information management. *Communications of the Association for Computing Machinery (CACM)*, 49(1):77–82, 2006.

18. J. Kelley and A. Chapanis. How professional persons keep their calendars: Implications for computerization. *The British Psychological Society*, 1982.

19. D. Kelly and J. Teevan. *Understanding What Works: Evaluating PIM Tools*, chapter 12, page 17. University of Washington Press, Seattle, Washington, May 2007.

20. J. Klingner, R. Kumar, and P. Hanrahan. Measuring the task-evoked pupillary response with a remote eye tracker. In *Eye Tracking Research and Applications Symposium*, Savannah, Georgia, 2008.

21. M. W. Lansdale. The psychology of personal information management. *Applied Ergonomics*, 19:55–66, 1988.

22. Y. Miyata and D. A. Norman. *The Control of Multiple Activities*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

23. D. A. Norman. *Emotional Design: Why We Love (or Hate) Everyday Things*. Basic Books, 2003.

24. R. D. O'Donnell and F. T. Eggemeier. *Workload assessment methodology*, chapter 2, pages 42/1–42/49. Handbook of perception and human performance: Vol. 2. Cognitive processes and performance. Wiley, New York, 1986.

25. A. Oulasvirta, S. Tamminen, V. Roto, and J. Kuorelahti. Interaction in 4-second bursts: The fragmented nature of attentional resources in Mobile HCI. In *CHI '05: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 919–928, New York, NY, USA, 2005. ACM Press.

26. S. Park, A. Harada, and H. Igarashi. Influences of personal preference on product usability. In *CHI '06: Extended Abstracts on Human Factors in Computing Systems*, pages 87–92, New York, NY, USA, 2006. ACM.

27. S. J. Payne. Understanding calendar use. *Human-Computer Interaction*, 8(2):83–100, 1993.

28. M. Pérez-Quiñones, M. Tungare, P. Pyla, and S. Harrison. Personal information ecosystems: Design concerns for net-enabled devices. In *Proceedings of the VI Latin American Web Congress - LA-Web 2008*, 2008.

29. M. Perry, K. O'Hara, A. Sellen, B. Brown, and R. Harper. Dealing with mobility: Understanding access anytime, anywhere. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 8(4):323–347, 2001.

30. P. Pyla, M. Tungare, J. Holman, and M. Pérez-Quiñones. Continuous user interfaces for seamless task migration. In *Proceedings of the 13th International Conference on Human-Computer Interaction, HCII 2009*, 2009.

31. P. Pyla, M. Tungare, and M. Pérez-Quiñones. Multiple user interfaces: Why consistency is not everything, and seamless task migration is key. In *Proceedings of the CHI 2006 Workshop on The Many Faces of Consistency in Cross-Platform Design.*, 2006.

32. J. Teevan, R. G. Capra, and M. Pérez-Quiñones. *How people find information*, chapter 3, page 17. University of Washington Press, Seattle, Washington, May 2007.

33. M. Tungare. *Mental Workload in Personal Information Management: Understanding PIM Practices Across Multiple Devices*. PhD thesis, Virginia Tech, 2009.

34. M. Tungare and M. Pérez-Quiñones. An exploratory study of personal calendar use. Technical report, Computing Research Repository (CoRR), 2008.

35. M. Tungare and M. Pérez-Quiñones. It's not what you have, but how you use it: Compromises in mobile device use. Technical report, Computing Research Repository (CoRR), 2008.

36. M. Tungare, P. Pyla, M. Sampat, and M. Pérez-Quiñones. Defragmenting information using the Syncables framework. In *Proceedings of the 2nd Invitational Workshop on Personal Information Management at SIGIR 2006.*, 2006.

37. M. Tungare, P. Pyla, M. Sampat, and M. Pérez-Quiñones. Syncables: A framework to support seamless data migration across multiple platforms. In *IEEE International Conference on Portable Information Devices (IEEE Portable)*, 2007.

38. S. Whittaker, Q. Jones, and L. Terveen. Contact management: identifying contacts to support long-term communication. In *CSCW '02: Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, pages 216–225, New York, NY, USA, 2002. ACM.

39. S. Whittaker and C. Sidner. Email overload: exploring personal information management of email. In *CHI '96: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 276–283, New York, NY, USA, 1996. ACM Press.