Available online at www.sciencedirect.com
ScienceDirect
Int. J. Human-Computer Studies 64 (2006) 941–961
International Journal of Human-Computer Studies
www.elsevier.com/locate/ijhcs

# Surviving task interruptions: Investigating the implications of long-term working memory theory

Antti Oulasvirta[a,*], Pertti Saariluoma[b]

[a]Helsinki Institute for Information Technology HIIT, Tammasaarenkatu 3, P.O. Box 9800, 02015 HUT, Finland
[b]Department of Computer Science, University of Jyväskylä, Finland

Received 12 August 2005; received in revised form 11 April 2006; accepted 21 April 2006
Communicated by D. Boehm-Davis

## Abstract

Typically, we have several tasks at hand, some of which are in interrupted state while others are being carried out. Most of the time, such interruptions are not disruptive to task performance. Based on the theory of Long-Term Working Memory (LTWM; Ericsson, K.A., Kintsch, W., 1995. Long-term working memory. Psychological Review, 102, 211–245), we posit that unless there are enough mental skills and resources to encode task representations to retrieval structures in long-term memory, the resulting memory traces will not enable reinstating the information, which can lead to memory losses. However, once encoded to LTWM, they are virtually safeguarded. Implications of the theory were tested in a series of experiments in which the reading of an expository text was interrupted by a 30-s interactive task, after which the reading was continued. The results convey the remarkably robust nature of skilled memory—when LTWM encoding speed is fast enough for the task-processing imposed by the interface, interruptions have no effect on memory, regardless of their pacing, intensity, or difficulty. In the final experiment where presentation time in the main task was notably speeded up to match the limits of encoding speed, interruptions did hamper memory. Based on the results and the theory, we argue that auditory rehearsal or time-based retrieval cues were not utilized in surviving interruptions and that they are in general weaker strategies for surviving interruptions in complex cognitive tasks. We conclude the paper by suggesting three ways to support interruption tolerance by the means of task and interface design: (1) actively facilitating the development of memory skills, (2) matching encoding speed to task processing demands, and (3) supporting encoding-retrieval symmetry.

*Keywords:* Interruptions; Working memory; Expertise; Human factors; Human–computer interaction; User psychology; Long-term working memory

## 1. Introduction

It is intriguing to note that in many, if not most, everyday situations we are actually able to overcome interruptions without notable costs to performance. Indeed, we almost always have numerous unfinished, simultaneous, successive, and overlapping tasks. When driving a car, we may need to keep in mind route information, speak to mobile phone about a family affair, visit restaurant, order something, return to the driving task, think about holiday, and keep all these tasks active until they are finished. Moreover, in switching from one task to another, people show remarkable ability in rapidly reactivating and accessing selected portions of their knowledge bases in a reliable and accurate manner. Skilled programmers, for example, are able to construct and keep updated large representations of code presented to them in a piecemeal and unorganized fashion. It is striking that this kind of *interruption tolerance* is possible at all, keeping in mind the known limitations of human working memory, and clearly, there are work domains where interruptions pose a serious problem to task performance. For example, after being interrupted, some workers do not resume their primary task 41% of the time (O'Conaill and Frohlich, 1995); and it took some office workers an average of 25 min

---

*Corresponding author. Tel.: +358 50 3841561; fax: +358 9 6949768.
*E-mail address:* antti.oulasvirta@hiit.fi (A. Oulasvirta).

1071-5819/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijhcs.2006.04.006

to return to their task (Mark et al., 2005). Understanding the circumstances that distinguish disruptive from non-disruptive interruptions is the general goal of the present paper.

We approach the problem from the perspective of memory. If one is to survive an interruption without substantial costs to memory, task representations must be stored so that they can be later accessed rapidly, reliably, and essentially in an error-free manner. We call such outcome of cognitive task processing the *safeguarding* of task representations. In this paper, we examine how safeguarding takes place; in other words, how people are able to keep processed task information in a state which allows overcoming interruptions without adverse effects. Particularly, we compare three alternative strategies people might use: maintenance through rehearsal during the interrupting task (or just before it), retrieval through temporal cues, and incidental encoding to cognitive structures called retrieval structures.

A satisfactory explanation to the above should target individual and task differences in interruption tolerance, as well as suggest how task environments and interfaces are to be organized to avoid interruption costs. We believe it is essential that research aims to understand the general characteristics of typical interruption situations related to concrete technological or task environments. The focus should be kept on characteristic interruption situations and the roles that various psychological constructs may have there. The ultimate goal of such research is to reach a comprehensive understanding of the interrupted mind in complex cognitive tasks rather than to achieve immediate improvements in particular technologies or tasks. Along these lines, we propose a theoretical approach to safeguarding, and investigate its nature in a series of experiments. In the end of the paper, interruption research is revisited from the perspective of the theory and, finally, implications to the design of tasks and interactive environments are considered.

## 2. LTWM and safeguarding

We use the theory of long-term working memory (LTWM for short; Ericsson and Delaney, 1999; Ericsson and Kintsch, 1995, 2000; Ericsson and Lehmann, 1996) to chart conditions for safeguarding in different situations. We believe that the theory is particularly suitable for addressing this practical problem, because its argumentation is based on observations and experiments made on skilled performance in cognitively complex activities. Ericsson and Kintsch (1995) argue that the performance of untrained subjects who memorize lists of unrelated items in the laboratory does not accurately describe the efficient storage and retrieval that people in specific domains can achieve after many years of practice. Moreover, only naïve subjects' performance is well described by standard models that have been developed to describe memory for unrelated materials in laboratory tasks (Ericsson and Polson, 1988).

### 2.1. Long-term working memory theory

In numerous experiments and domains, it has been shown that experts are able to store very large representations of task-specific information, quicker and more accurately than novices. Chess researchers observed this as early as in the 1920s (Djakov et al., 1926). This was also associated with cognitive skills by de Groot (1965, 1966; Vicente and de Groot, 1990). Expert knowledge's meaning and connection to working memory was illustrated and theoretically analysed by Chase and Simon (1973). Shneiderman (1976) found experts' memory superiority in computer programming, and Sloboda (1976) in music (see also McKeithen et al., 1981; Vicente, 1988). Similarly, Go and bridge experts (Charness, 1979; Engle and Bukstel, 1978; Reitman, 1976), electronics engineers (Egan and Schwartz, 1979), trained map-readers (Gilhooly et al., 1988; Kinnear and Wood, 1987), digit-memorizing specialists (Chase and Ericsson, 1982; Ericsson and Staszewski, 1989), expert waiters (Ericsson and Polson, 1988) and radiologists (Mylers-Worsley et al., 1988), expert architects (Akin, 1986) and taxi drivers (Kalakoski and Saariluoma, 2001) recall task-related stimuli remarkably better than novices, presumably because of their large stores of prelearned chunks of information (Charness, 1988, 1989, 1992; Ericsson and Charness, 1994; Ericsson and Lehmann, 1996; Green and Gilhooly, 1992; Saariluoma, 1985, 1989, 1995; Saariluoma and Kalakoski, 1997, 1998).

In addition to being able to store larger representations, experts were also found to be better able to retain them over intervals of interrupting activities. A great deal of literature in expert memory research suggests that interrupting tasks do not necessarily have a very strong negative effect on memory traces of primary tasks (Charness, 1976; Edwards and Gronlund, 1998; Fischer and Glanzer, 1986; Frey and Adesman, 1976; Gillie and Broadbent, 1989; Glanzer et al., 1981, 1984; Lane and Robertson, 1979). Moreover, information-processing demands of the interrupting task have rarely effect on experts. For example, a study by Ericsson and Polson (1988) showed that an expert waiter was able, almost without an error, to remember orders from a whole day in a post-session free recall test. Moreover, experts are less disrupted, systematically, than novices by interruptions during high memory load. For example, players of blindfold chess must remember up to 10 games simultaneously without seeing the boards (Saariluoma, 1991), and this performance is hampered by concurrent but *not* by posterior interference by another imagery task.

These and other findings have questioned the notion that all task-relevant information is held in short-term memory (STM) (e.g., Chase and Simon, 1973), and suggested a more active role for long-term memory (LTM) in storing task representations. On the grounds of experts' interruption tolerance, a very fundamental change in the working memory theory has recently taken place. Instead of assuming that working memory is a unified store, it has become evident that it has, in addition to the short-term part (STWM), a long-term part in LTM called the *Long-Term Working Memory* (Ericsson and Kintsch, 1995; see also, Baddeley, 2000; Gobet, 1998; Richman et al., 1995; cf. Vicente and Wang, 1998). In many skilled activities, the long-term part must be utilized to overcome the severely limited capacity of STWM. As known already from the work of Ebbinghaus (1885), when chunks are stored to LTM, it provides a practically unlimited time and capacity for storage. By contrast, information held in STWM is known to be in a fragile state, vulnerable to almost any disruption in processing and maintenance (e.g., Baddeley, 1986; Cowan, 2001).

The key postulate in explaining how LTM is utilized is called *retrieval structure*. Retrieval structures are abstract, hierarchical knowledge representations that experts develop to efficiently encode and retrieve large amounts of information in LTM. Retrieval structures consist of two types of associations: First, associations to *a system of cues are* repeatedly extracted to index-specific semantic categories of information, the function of which is to "allow retrieval of the most recently encoded information through reinstatement of those retrieval cues, even when this type of information is frequently updated and changed during the processing" (Ericsson and Delaney, 1999, p. 579). Secondly, these cues are embedded in *generated structures* in LTM, where information presented is interrelated to other pieces of presented or generated information, as in the classic case of expert digit recall performance (Ericsson and Staszewski, 1989). The idea of *slots* in *generic* retrieval structures, "that individual elements can be stored independently of meaningful associations of other elements at fixed time durations", has been rejected in the theory (Ericsson and Kintsch, 2000, p. 572).

LTWM is distinguished from STWM by the durability of the storage it provides and by the need for sufficient retrieval cues in attention for access to information (Ericsson and Kintsch, 1995). When elements of retrieval structures remain activated in LTWM, an access to the information is retained, although it is indirect. When forced to interrupt performance, activation decreases, causing difficulties in using temporal recency as a retrieval cue, but the structures remain intact in LTM. Upon resuming, reinstatement of the relevant structures takes place by first initiating the task representation and related strategies and then accessing associated structures (Ericsson and Delaney, 1999).

### 2.2. *On the nature of LTWM encoding*

Efficient use of retrieval structures enables *distinctiveness* of the encoded materials, meaning easier distinction of the encoding from other task representations and from pre-existing semantic knowledge upon recall. *Selectivity* of encoding (favoring of some features or pieces of information while neglecting others) also increases with practice in accord with task's retrieval demands. Moreover, these skills allow *anticipatory* encoding of retrieval cues: When subjects have experience with the task demands and have acquired stable procedures for completing the task, they can foresee retrieval demands and develop memory skills to index relevant information to task representations (Ericsson and Kintsch, 1995). Taken together, the use of many types of cues (e.g., categorical and temporal) in a manner anticipating what cues are available upon retrieval, combined with traversal of the hierarchical structure to unfold the needed information in a correct order, allows for overcoming proactive interference caused by prior storage of similar information in LTM.

The theory of LTWM rejects the notion that LTM corresponds to a unitary process that generalizes to all types of materials and requires exactly the same amount of time for successful storage. Therefore, efficient use of LTWM is not possible in all circumstances but vitally relies on memory skills acquired in particular domains and activities through practice and training. In the following, we summarize some practically important *conditions and characteristics of safeguarding*. These are adopted from the literature on Skilled Memory Theory (LTWM's predecessor) and LTWM (Chase and Ericsson, 1982; Ericsson and Delaney, 1999; Ericsson and Kintsch, 1995; Ericsson and Lehmann, 1996; Ericsson and Polson, 1988; Ericsson and Staszewski, 1989).

(1) *Practice-Dependence.* The development of LTWM is integrally tied to consistent and deliberate practice in the task domain (see Ericsson and Lehmann, 1996, for qualities of good practice and training in different domains). In addition to consistency and deliberation, another important quality of practice is *memory-based selectivity*, the requirement that the task involves selecting between alternatives that are not directly perceivable but are kept in memory. Tasks that can be performed primarily based on immediately available perceptual information (e.g., graphical user interfaces with WYSIWYG) place lower pressures on the building up of the necessary memory skills for LTWM (e.g., than command languages). Even apparently perceptual tasks such as chess, however, may require simulating events in the mind's eye, which demands keeping in memory large representations. (Ericsson and Kintsch, 1995, cf. Hatta et al., 1989.)

(2) *Speed up.* The speed up principle states that with practice, normal subjects can increase the speed of storage and retrieval in LTM so that it approaches the speed of information storage and retrieval in STM. This is because of the contribution of LTM in chunking and elaboration of materials. On the other hand, STWM's restricted capacity in encoding poses limits to how fast complex information can be encoded. A significant amount of empirical evidence from various domains—mainly chess, mental calculation, reading, and expert waitresses—has been found to support this principle (for a review, see Ericsson and Lehmann,

1996). Moreover, this effect—closely related to the practice effect and often regarded as theoretically unimportant in experimental psychology (Ericsson and Delaney, 1999)—has been observed with only small amounts of practice. For example, Edwards and Gronlund (1998) found interrupted memory performance analogous to LTWM in just 6 experimental trials with a fixed presentation order (see also Detweiler et al., 1994; Trafton et al., 2003).

(3) *Meaningfulness*. Orientation towards to-be-encoded materials is *meaningful* and accords with a vast body of prior knowledge. This is well demonstrated in the levels-of-processing paradigm (Craik and Lockhart, 1972; Hyde and Jenkins, 1969): When subjects are presented with words and required to respond to an orienting question (e.g., shallow orientation: "Is this word in uppercase?" or deep orientation: "Does this word refer to an animal?"), and memory is subsequently tested, it has been consistently found out that memory is strongest for words subjected to a deep (or meaningful) orienting task (see also Oulasvirta, 2004, for a demonstration of this effect in HCI). However, by devoting more and more time to encoding, and with deliberate practice, even arbitrary materials such as random digit series can be encoded (Ericsson and Staszewski, 1989).
(4) *Organization.* Task processing organizes the materials into a hierarchically organized system of retrieval cues, the retrieval structure (explained above). This also means that there must be such constraints in the materials so that they can be organized (see Vicente and Wang, 1998).
(5) *Incidental encoding.* Consequently, information is incidentally encoded during skilled activities, as outcomes and intermediate products of complex cognitive processing of the task, to LTWM.
(6) *Domain-specificity.* Finally, the resulting skills, procedures, and structures in memory are restricted to the domain or activity of practice. "Experts use their knowledge structures in semantic memory to store information during skilled performance of some task" (Chase and Ericsson, 1982, p. 159).

### 2.3. *Objectives of the study*

In this paper, we are interested in critically testing the implications of the LTWM theory in areas of investigation of interest to readers of IJHCS. Particularly, we are interested in information workers' ability to perform an accurate, reliable, and practically error-free regeneration of an interrupted task. According to the theory, people are able, in restricted conditions (see Section 2.2.), to retain a high level of memory performance over interruptions of relatively long duration. If safeguarding to LTWM has taken place, the resumption of a task should be quick and accurate regardless of the processing demands of the interrupting task. However, when the demands of the main task are at odds with memory skills, an interruption cost should emerge, showing a clear deterioration of memory performance. This proposition is evaluated in the set of experiments reported here.

In addition to investigating the *existence* of interruption tolerance, we are interested in critically evaluating the LTWM theory in explaining *how* the safeguarding of information takes place in a complex cognitive task. The unique claim of LTWM is that the capacity of working memory can be extended by efficient use of LTM, particularly so that there is no need for maintaining information in STWM, but LTM can be used for reinstating the task representations after an interruption (see above). This efficient utilization of LTM underlies people's ability to survive intense and long interruptions in their skilled activities, without utilizing special strategies like rehearsal. In this paper, two alternative accounts, both used to explain memory in interrupted activities, are contrasted to LTWM. Generally speaking, both assume a decrease in memory performance in conditions where LTWM would not.

Firstly, within current theories of STWM, short-term traces are assumed to decay completely, in a time of about 30 s, when no rehearsal can be carried out during the interruption (e.g., Cowan, 2001; Baddeley, 1986). Therefore, some *strategy* of retaining information in STWM is needed. The first competing hypothesis to explain how LTM can be employed during task performance is that we store a *tag* of the interrupted task in our STWM instead of the chunk itself (Gobet and Simon, 1996a, b, 1998; Gobet, 1998; Simon, 1976). This tag, or *pointer*, takes only the space of one chunk in the STWM whereas the referred-to task in long-term storage can contain several chunks (Gobet, 2000a, b, 2001, see also Ericsson and Kintsch, 2000). The idea is that unless the interpolating activity is very intense, the tag can be retained in STWM over the interruption and used for instant resumption of the task. This assumes that there is some "special strategy" that is used to periodically refresh the tags in STWM during an interruption. In our experiments, we test this idea by creating very demanding interruption conditions that would hamper maintenance of tags over the interruption.

The second main alternative (Schneider and Shiffrin, 1977; Anderson, 1983) proposes a uniform representation of memory, in which STM is tantamount to *activated* elements in LTM. According to Ericsson and Kintsch (1995), theories of working memory based on *transient activation of information in LTM* cannot explain the resumption of activity once the information in working memory has been irretrievably lost. The information cannot be distinguished from all the other information in LTM, because it does not have a higher level of activation. Activation-based models have been used to analyse interruptions (see e.g., Trafton et al.'s, 2003, ACT-R based goal-activation account). In such models, successful retrieval over a long period of interruption critically depends on activation built up either during the main task or during the interruption lag (the interval between an alert to pending interruption used for rehearsal to build up activation). While it is impossible to determine exact parameters for these models so that they could be utilized to model our particular experimental situation, these models generally predict a significant loss of activation and thus compromised memory performance, given that the duration of interpolated activity is as long as 30 s as in our experiments. In addition, they generally predict that interruption would not directly cause omissions, as the tag hypothesis suggests, but *proactive interference*, which in turn would be manifested in a decreased ability to make a difference between the representation of the main task and related pre-existing knowledge in LTM. This should result in both omissions and false recognitions. We approach testing this idea by distinguishing between these two types of memory errors in the post-interruption memory test in our experiments.

## 3. Experiments

In this section, we describe four experiments designed to assess the hypotheses. We first describe the general method and then go through each individual experiment, reporting specific hypotheses, details of method, results, and discussion.

### 3.1. General method

To approach the predictions experimentally, we needed a task domain that would be important to IJHCS and where it would be easy to find skilled subjects. Although there are individual differences in the use of memory during reading, there are many reasons to suggest that *reading and comprehension* is a *task domain* where most educated adults are skilled in (Ericsson and Kintsch, 1995). We start practicing reading and comprehension at the age of six and seven and, as adults, we read papers, books, documents, web pages, magazines and watch and listen to news on a daily basis. Thus, most of us, and all of the participants of our experiments, have at least 10 years of experience with this type of information processing (coinciding with "the 10-year rule of expertise" of Chase and Simon, 1973). Moreover, because news and other expository texts are serially presented and often talked about with other people, in the absence of the medium itself, special burden is placed to develop skills to use LTM efficiently. As the format or script of expository texts is usually consistent, our practice with the activity has taken place in such circumstances that allow for the development of anticipatory and selective retrieval cue encoding. Finally, we assume that our subjects are interested in "doing well" in the experiment (due to evaluation apprehension and academic compensation for participation), they are also motivated to devote attention and effort to processing of the experimental materials.

As argued in the theory, the use of LTWM is dependent on the efficient use of existing knowledge in LTM. Citing three studies on this subject, Ericsson and Kintsch (1995) claim that pre-existing *domain knowledge*, rather than reading ability or IQ, influences memory and comprehension of texts. Thus, good readers, instead of having "more room" in STWM, are claimed to be able to construct larger and more integrated retrieval structures to LTWM during reading through their use of more sophisticated encoding procedures. Because of this domain-specificity, we needed to acquire such materials that their style and content would be familiar to the participants without being trivial to remember (and thus result in a ceiling effect). For this end, we chose to use *expository texts* of encyclopaedic style. The texts were sampled from the topics of History and Natural Sciences and involved titles such as "Anatomical Differences Between Man and Ape," "World's Oil Resources," "Global Climate Types," "Ayatollah Khomeini's Regiment in Iran," and "The French Revolution," all of which are studied in schools, presented regularly in news, documentary films, newspapers, magazines, radio, etc. We can assume that participants (university students) have had considerable amount of deliberate and consistent practice in the domains of the texts prior to the experiment.

The basic paradigm is as follows. We used two tasks in the present study. One is simply called *the main task* and the second *the interrupting task*. The main task consists of computer-paced, one-by-one reading of sentences. After 30 s, it is interrupted by the interrupting task, a multiplication verification task (hereafter: *MVT*), after which the primary task continues for another 30 s. In order to minimize the amount of retroactive interference caused by similarity of the interrupting task, which is known to hamper even skilled performance (e.g., Charness, 1976; Oulasvirta and Saariluoma, 2004), we selected the interrupting task (multiplication verification) from a domain completely different to the domain of the main task (reading). In the control condition, no interruption is presented. Memory is then tested in a sentence recognition test.

It is important to keep in mind that LTWM predicts differences across types of reading materials within the reading task. Interruption tolerance should span only *semantic levels of text representation* and not the *surface code*. Most psychologists studying text representation agree that there are five levels of text representation (Graesser et al., 1997; Kintsch, 1988). These five levels are (1) surface code, (2) the textbase, (3) the situational model, (4) the communication level, and (5) the genre. The surface code preserves the exact wording and syntax of the text. The textbase captures the meanings of the words in a propositional form, but ignores the exact working and syntax. The situation model (or mental model) is the non-linguistic, referential content of what the text is about. It may contain component hierarchies, causal chains, and spatial regions. The communication level contains the pragmatic context that frames the messages in the text (e.g., who is communicating with whom?). Finally, the genre level assigns the text to one or more rhetorical categories and uses the selected text genre (e.g., science textbook, novel, manual, document, report etc.) to guide comprehension. Importantly, only the last four levels are encoded steadily to LTWM, whereas the surface code is evidenced to be rapidly forgotten (see Ericsson and Kintsch, 1995; Kintsch, 1988). Therefore, to operationalize "semantic levels of text representation", we use a method for measuring memory for the gist by Royer et al. (1987) in which the surface features of all test items are different from those seen in the text, but the semantic content is different in half of the test items (and same in half of the items). Participants are required to judge which propositions were made in the original text and which not. In order to distinguish omissions from false recognitions, an "I don't know" option is provided as well.

### 3.1.1. Participants

A total of 108 subjects contributed data to the experiments reported here. The subjects are undergraduate and Master's students recruited from cognitive science courses of two universities. Their ages ranged from 19 to 37 years with an average of 23 (SD = 3.9). Sixty-four percent of the participants were male. They received partial course credit or money for participation. Details of recruitment are reported in association with each experiment. All subjects were native speakers of Finnish, the language of the experimental materials.

### 3.1.2. Materials

Two sets of texts consisting of 12 sentences were used in the experiments. The first set consisted of texts related to *natural sciences* sampled from an encyclopedia (Uusi Pikkujättiläinen, WSOY, Finland, 1984). The topics were: "Finnish Swamps," "Early Years of Space Research," "Anatomical Differences Between Man and Ape," "World's Oil Resources," "Global Climate Types," and "Finnish National Parks." Each text consisted of a header and 11 sentences, each sentence spanning 8 to 11 words. The second set concerned *history*. The topics were: "Ayatollah Khomeini's Regiment in Iran," "The French Revolution," "The Plague and the Persecution of Jews," "Dissertation Practices in the Middle Ages," "Scientology and L. Ron Hubbard," and "Pyramid Robbers in Ancient Egypt."

*Multiplication verification* was used as the interrupting task. The method was adopted from Byrne and Anderson (1999) who had shown that it is sufficiently engaging to disturb dual-task performance. A speeded response was to be given on whether a presented equation was correct, for example, "Is $8 \times 8 = 64$ correct?" Half of the equations were correct, half incorrect.

All questions in the sentence recognition test were of the form "Was it claimed in the text: X." Six of the items concerned the passage (headline and first five sentences) read before the interruption, and six of them concerned the passage (six last sentences) read after it. Half of the claims were correct and half incorrect. To reduce guessing artifacts, and to distinguish omissions from intrusions, "I don't know" answers were permitted (Gardiner and Richardson-Klavehn, 2000). Test items that required rejection were constructed by sampling common claims related to the topic from the Internet and encyclopedia.

The *to-be-rejected* test items were intentionally made intrusive by making them semantically related to the passage and familiar (well known to the public, and possibly true, as judged by the experimenter). This manipulation has been shown to be effective for creating claims that are likely to intrude and create false memories (Braun et al., 2002). The method used here for creating the *to-be-accepted* test items is the *Sentence Verification Technique* by Royer et al. (1987). In constructing paraphrasing sentences, Royer et al. suggest taking the to-be-recognized sentence as the basis and then changing its words for their synonyms, changing the ordering of words, and finally making the new sentence shorter or longer by adding or removing words. This method was used to reduce the possibility of recognizing a sentence from verbatim or visual memory (i.e., surface code memory) and to increase the demand for using semantic memory.

All materials were in Finnish.

### 3.1.3. Procedure

Subjects were told that they would be participating in an experiment studying the effects of mental exertion on memory. Use of response keys, multiplication verification task, reading task, and recognition task were practiced separately. All responses were asked to be made as quickly and as accurately as possible. All subjects were informed that when responding to the recognition test they *should not confound what they already know about an issue with what was said about it in the text.* Hence, they were in effect warned against false recognitions.

Presentation times of sentences in the main task are reported below individually for each experiment. In all experiments, the interrupting task occurred after the first six sentences. After exactly 30 s of doing the MVTs, the reading task continued with the six remaining sentences from the text. In the control condition, reading was *un*interrupted, except in Experiment 3 where easy and hard interruptions were compared. The sentence recognition test was self-paced and comprised 12 questions. Six test items for the first half of the texts were presented in randomized order before test items of the second half. After responding to a test item, subjects had to indicate their confidence for their response by pressing a number key (from 1 to 7, with 7 signifying most confidence). After each trial, a possibility to rest was provided after each test.

Stimulus presentation was controlled by a PC with millisecond accurate timing. Sentences were shown in the middle of a 17" display with a 22-point font. All responses were given on a Finnish QWERTY-keyboard. Key "-" was designated as "Yes," "<" as "No," and spacebar as "Don't Know."

### 3.1.4. Measurements

The *effect of interruption* is calculated separately for correct recognitions, omissions ("I don't know" answers), and false recognitions by subtracting responses obtained in the *un*interrupted condition from those obtained in the interrupted condition. If interruptions have a negative effect on memory, one should consequently find a negative interruption effect score on correct recognitions and/or a positive score for false recognitions and omissions.

The results pertaining to subjective confidence levels are omitted from this paper for the sake of simplicity. However, it is worth mentioning that the confidence level data did not contrast the accuracy data in important respects.

To assess if memory performance was above chance level in an experiment, a *total memory accuracy* measure is calculated by subtracting false recognitions from correct recognitions. This measure gives an average score of zero for three types of random answer behaviours: (1) answering yes or no to all test items, (2) answering "I don't know" to all items, and (3) randomly selecting yes, no, or "I don't know" to all items. (Recall that participants were asked to avoid false recognitions.)

### 3.1.5. Statistical analysis

The default statistical test employed throughout the experiments for the four measures is a single sample $t$ test comparing the effect of interruption to zero, except for Experiment 3 where ANOVA is used. An alpha-level of .05 is utilized throughout.

## 3.2. Experiment 1

Generally speaking, we wanted to test the hypothesis of LTWM according to which people are able to safeguard task information in their skilled activities. If safeguarding is possible in a computer-paced task like this, even a relatively long interruption should cause no decline in their memory performance.

### 3.2.1. Method

*3.2.1.1. Participants.* Twenty-four students were recruited for the experiment; 12 students from an introductory course in experimental psychology at the University of Helsinki and 12 similar students from a course on HCI at the University of Jyväskylä. Both groups received a partial course credit for participation.

*3.2.1.2. Materials.* The natural sciences text set was used in this experiment. Multipliers in the interrupting task ranged from 4 to 9.

*3.2.1.3. Design.* The reading condition variable, interrupted vs. continuous reading, was controlled within subjects. Each subject completed six trials (in addition to one practice trial). Counterbalancing was accomplished by rotating texts and experimental conditions across subjects, yielding a total of 12 combinations.

*3.2.1.4. Procedure.* Sentence presentation time in the main task was 6000 ms. A new MVT was presented every 3000 ms.

### 3.2.2. Results

We here present the statistical analyses; data on interruption effects are presented in Table 1. (For raw memory scores, see Appendix A.)

A single sample $t$ test for total memory accuracy (against 0) was statistically significant, $t(23) = 11.60$, indicating that the level of performance was above chance level.

Effects of interruptions on the three memory performance scores in Experiments 1, 2, and 4 are reported in Table 1.

In this experiment, interruption had no effect on correct recognitions, $t(23) = .00$, on false recognitions, $t(23) = .18$, nor on omissions (i.e., "Don't know" answers), $t(23) = -1.23$.

### 3.2.3. Discussion

To sum up, Experiment 1 shows no effects of interruption whatsoever on memory accuracy in computer-paced reading of an expository text. The results thus corroborate the hypothesis that the upkeeping of task representations in this activity can be robust to interruptions even though the interruption task is quite demanding and as long as 30 s in duration.

By contrast, the results are not in line with the tag hypothesis, the view according to which the locus of task representation storage is in STWM, not LTM. An interruption that captures attention for 30 s should be long enough to seriously hamper the upkeeping of the tags in STWM, and lead to difficulties in reinstating the correct representation upon task resumption, which should show up as memory losses.

Table 1
Effects of interruptions on memory performance in different encoding conditions (Experiments 1, 2, and 4)

| Experiment | Presentation speed (ms) | | Interruption effect[a] | | |
|---|---|---|---|---|---|
| | Main task | Interruption | Correct recognitions | False recognitions | Omissions |
| 1 | 6000 | 3000 | .00<br>(3.60) | .20<br>(2.86) | −.62<br>(2.48) |
| 2 | 6000 | Self-paced | .13<br>(4.81) | −.63<br>(3.91) | .38<br>(2.56) |
| | 6000 | 3000 | 1.63<br>(4.36) | −1.38<br>(3.16) | −.25<br>(3.19) |
| 4 | 2500 | 3000 | −1.54*<br>(3.48) | 0.04<br>(3.13) | 1.42<br>(3.49) |

*Note*: Numbers in parentheses denote standard deviations.
*$p < .05$.
[a] Calculated by subtracting scores in the uninterrupted condition from scores in the interrupted condition.

The results do not provide direct support to the activation-based models of interruptions either. There, as argued above, substantial deactivation of information should occur over an intervening period as long as 30 s, leading to omissions and possibly interference.

However, both alternative hypotheses can still be entertained if the interruption task was not intensive enough in our experiment but allowed periodical refreshing of the tags, or building activation, with some special strategy while answering to the MVTs. We turn to this possibility in the next experiment.

### *3.3. Experiment 2*

We were worried that the interrupting task in Experiment 1 might not have been intensive enough. It might have allowed the participants to engage in active refreshing, perhaps by the means of inner speech, during the 30-s interrupting task. Several studies have shown that the articulatory loop can be used to remind oneself of the to-be-switched to task in an alternative tasks paradigm (Baddeley et al., 2001; Emerson and Miyake, 2003), and this kind of strategy has been proposed to explain also experts' ability to survive interruptions (Gobet, 1998). Hence, we needed a situation where the interruption task would be so intense that the possibility for rehearsal was minimal. The 3000 ISI of the interrupting task might have left just enough time after answering the MVT for rehearsal in STWM.

In Experiment 2, half of the subjects were instructed to do the interrupting task as quickly as possible (*self-paced* group), whereas the rest were given an experimenter-paced task with an ISI of 3000 ms (*experimenter-paced* group, as in Experiment 1). We expect the self-paced group to perform more MVTs and thus have fewer opportunities to rehearse. If information is indeed kept active by intermittent rehearsal, minimizing the possibility to strategically maintaining them by enforcing a higher intensity of interrupting activity should hamper remembering the first part of the task and thus hamper task resumption. As hypothesized in Experiment 1, this should be manifested in increased omissions and false recognitions due to the interruption. While the experimenter-paced group should basically replicate the null effect of Experiment 1, the self-paced group should show a clear interruption effect. Should there be such a disadvantage for the self-paced group, it would be interesting to know how exactly the maintenance of the tags was scheduled. We therefore analyse reaction times for the MVTs.

On the other hand, if LTWM is driving the maintenance of task representations, the processing demands of the interrupting task should not matter, which would imply another null effect also for the self-paced group. From this perspective, Experiment 2 is predicted to replicate the behavioural results of Experiment 1 with different main task materials (a history text set instead of a set of natural sciences texts).

#### *3.3.1. Method*

*3.3.1.1. Participants.* Twenty-four students of cognitive science were recruited from the University of Helsinki. A small (10 €) incentive was paid for participation.

*3.3.1.2. Materials.* The history text set was used in this experiment. The interruption task was also identical to Experiment 1.

*3.3.1.3. Design.* The reading condition was controlled within subjects on two levels, interrupted vs. continuous reading. Each subject completed six trials (in addition to one practice trial). Counterbalancing was accomplished by rotating texts and experimental conditions across subjects, yielding a total of 12 combinations.

Pacing was controlled between subjects because changes in pacing within subjects were considered confusing by a subject in a pilot study.

*3.3.1.4. Procedure.* The procedure was identical to that in Experiment 1, except for changes due to the introduction of the between-subjects variable. In the experimenter-paced group, a response to the presented equation was to be given during 3000 ms, whereas in the self-paced group, a new equation was immediately presented after each response.

After 30 s, reading continued with the six remaining sentences from the same text. In the control condition, reading was uninterrupted.

#### *3.3.2. Results*

A single sample $t$ test for the total memory accuracy score was statistically significant, $t(23) = 19.21$, indicating that the level of performance was above chance level.

Interruption effects in the two groups are here reported both separately and combined. There were no effects of interruption on correct recognitions in neither group (both $\mathrm{df} = 11$) nor over all subjects combined ($\mathrm{df} = 23$), all $|t\mathrm{s}| < 1.29$, all $p\mathrm{s} > .22$. There were no effects of interruption on false recognitions in the in neither group nor in the combined score, all $|t\mathrm{s}| < 1.51$, all $p\mathrm{s} > .16$. Finally, there were no effects of interruption on omissions in neither group nor in the combined score, all $|t\mathrm{s}| < .51$, all $p\mathrm{s} > .62$.

In order to assess if subjects in the self-paced group really performed more MVTs, MVT performance was analysed for both groups. In the self-paced group, at least 17 multiplications were correctly verified in 50% of the trials during the given 30 s, the maximum being 30. Hence, significantly more MVTs were done in the self-paced group than in the experimenter-paced group (10). Interestingly, the difference in reaction time between the first ($M = 3157$ ms) and the second ($M = 1807$ ms) multiplication equation was 1350 ms. In the experimenter-paced group, responding to the first MVT took about 1150 ms less than in the self-paced group, but after that the self-paced group was much faster in responding.

#### 3.3.3. Discussion

When we look at the data combined over the two groups, we can see that the null results from Experiment 1 were replicated with different materials.

In line with the prediction of LTWM, the data shows no clear disadvantage or advantage for having to perform the interrupting task at the own maximum pace, although the analysis of responses in the interruption task revealed that as much as 70% more MVTs were done by the self-paced group than the experimenter-paced group. This means that significantly less time could have been devoted to the strategic rehearsal during the interruption task. These observations are at odds with the tag hypothesis, according to which intensive engagement in another task should wipe out the tags in STWM that would be needed for quick and error-free reactivation of the task representation upon task resumption.

The results are at odds with activation-based models as well, but only if no rehearsing (to build up activation) took place just before switching to the interrupting task. For example, according to Trafton et al.'s (2003) model, activation of the main task can be built by rehearsing it during or just before the interruption task. An interesting observation relating to this was in the analysis of MVT RTs: responding to the first stimulus in the interruption task took about 1150 ms longer in the self-paced group than in the experimenter-paced group, but immediately after this, the self-paced group performed in a much higher pace. Could this mean that the self-paced group used this interruption lag to rehearse the main task? We believe there are reasons suggesting that *not* enough rehearsal could take place in such a short period. The rehearsal rate of phonological loop is inherently too slow for rehearsing a text consisting of six sentences, or a gist/summary of it, in just 1 s (Zhang and Simon, 1985; Baddeley, 1986). Moreover, the times needed for efficient rehearsing during an interruption lag are typically much longer. For example, Trafton et al. (2003) showed advantages for an interruption lag of as long as 8 s.

### 3.4. Experiment 3

In the third experiment, we sought to investigate the possibility that rehearsal would be used as a strategy when memory performance is low in the main task. In the previous experiments, because of the relatively long presentation time in the main task, subjects would have had time to build up enough activation to overcome the interruption, and would not have needed to engage in effortful strategic rehearsal. We approached this question by creating a condition in the main task where memory performance is necessarily very poor and then, in first condition, provide a possibility for engaging in rehearsal and, in the second condition, make the interruption so difficult that no rehearsal is possible.

The experimental task is similar to Experiment 1, although the control condition of uninterrupted reading is not included in this experiment. The essential manipulation here is that of *orientation to text*. We instructed how sentences were to be read in each reading condition. Participants either rehearsed the sentence as quickly as possible, made a judgment on the sentence being easily imaginable, or decided whether the sentence had exactly two "p" letters (they occur often in Finnish words). An easy or hard interruption task occurred unexpectedly halfway through reading. The easy interruption had an inter-stimulus interval of 5000 ms, whereas in the hard task it was 2000 ms. In the easy interruption, multipliers ranged from 1 to 4, whereas in the hard interruption they ranged from 6 to 9, thus making the hard interruption significantly more taxing. This manipulation of difficulty was adopted from Byrne and Anderson (1999), who showed that the two differ in their attentional demands in dual-task situations.

Given that processing time per sentence is kept constant across the three orientation conditions, scores across the three orientation conditions should reflect the Craik and Lockhart's (1972) classic levels-of-processing effect. In the letter-search condition, a weak overall level of memory performance is expected because of the orientation towards the surface features. Better performance is expected in the two other conditions, the imagination condition entailing the best overall level. No activation-based models of interruption published thus far have attempted to model how *orientation* in the main task affects interruption tolerance, although the more general models could allow for this (Bradshaw and Anderson, 1982, modeled the levels-of-processing effect with ACT-R). However, it is reasonable to suggest, in the context of activation- and tag-based models, that the letter-search condition would result in a weaker representation, which could motivate the subjects to engage in strategic rehearsal as a way to maintain this information.

The rehearsal condition is of interest here as well. The use of rehearsal already in the main task could make it easier to *continue* using that strategy in the interruption. In general, if intermittent rehearsal is adopted as a strategy to compensate for low level of performance in the main task, superior accuracy should be observed in the easy interruption condition that provides better resources for rehearsal. The tag model assumes a general adverse effect to memory performance in the hard interruption condition because of decreased possibilities for intermittent rehearsal. This would be manifested in better memory performance in the easy condition across the three encoding conditions.

By contrast, the LTWM theory suggests that the complexity or difficulty of the interrupting task should not matter even in poor encoding conditions. The difficulty of interruption should have no influence on the measures of recognition accuracy, but yet another null effect should be observed.

#### 3.4.1. Method

*3.4.1.1. Participants.* Forty-one students from an introductory cognitive science course at the University of

Helsinki and from a similar course at the Open University of Helsinki participated in this experiment in partial fulfillment of course requirements. Five subjects were excluded from the analysis due to their inability to follow instructions (i.e. constantly forgetting to perform the level-of-processing task that was instructed, or misunderstanding the instructions).

*3.4.1.2. Materials.* Texts and recognition test items were adopted from Experiment 1.

As in Experiment 1, the interrupting task required subjects to verify presented multiplications. In the *easy condition*, multipliers ranged from 1 to 4, whereas in the *hard condition* they ranged from 6 to 9.

*3.4.1.3. Design.* There were two independent variables, both within subjects. The first variable, *level-of-processing*, had three levels: imaginary-task, letter-pair search, and rehearsal. The second variable, *interruption-difficulty*, was controlled on two levels: easy and hard.

Each subject completed six trials each (in addition to the practice trials described in the Procedure section). Counterbalancing was accomplished by rotating texts, interruptions, and levels-of-processing conditions across subjects.

*3.4.1.4. Procedure.* Instructions for each level-of-processing manipulation were given on the screen and explained verbally by the experimenter prior to each trial. Instructions for the different level-of-processing-tasks were as follows: "Your task is to judge whether the state of affairs described in the sentence can be imagined or not," (imaginary task) "Your task is to judge whether there are *exactly two* p-letters present in the sentence," (letter-pair search) and "Your task is to rehearse the sentence in your mind and press 'yes' each time you have read the sentence, word-by-word, from the beginning to the end" (rehearsal). The rationale for asking subjects to search for *exactly* two p-letters is to ensure that sentences are read in their entirety. Each sentence was presented for 6000 ms, during which the yes/no response was to be evoked. The easy conditions differed from the hard ones in their inter-stimulus interval of the MVTs (2000 vs. 5000 ms, respectively). After 30 s of the interruption task, the reading task continued with the last block of sentences. The recognition test was self-paced and comprised 12 questions. In contrast to Experiments 1 and 2, confidence levels were not asked for. After each trial, subjects had a possibility to rest.

### *3.4.2. Results*

A single sample *t* test for the total memory accuracy score was statistically significant in all encoding conditions, all $ts(35) > 6.97$, all $ps < 0.01$, indicating that the levels of memory performance were above chance level. A one-way repeated measures ANOVA for total memory accuracy revealed a significant effect of encoding (reading) condition, $F(2, 70) = 5.74$, $p < 0.01$. As expected, Imagine condition was better than Letter-Pair Detect, which in turn was better than Rehearsal. This trend replicates the Levels-of-Processing effect (Craik and Lockhart, 1972).

Effects of interruption-difficulty on the three accuracy measures in the three encoding conditions are presented in Table 2. There were no effects of interruption-difficulty (as calculated by subtracting scores in the easy interruption condition from the scores in the hard condition) on correct recognitions in any of the encoding conditions nor in their combined score, all $|ts(35)| < 1.49$. There were no effects of interruption-difficulty on false recognitions in any of the encoding conditions, or in their combined score, all $|ts(35)| < 1.56$. Finally, there were no effects of interruption-difficulty on omissions in any of the encoding conditions, or in their combined score, all $|ts(35)| < 1.76$, all $ps > .08$. (The effect of interruption was most prominent, although not very close to being statistically significant, on omissions in the Letter-Pair-Search condition, $t(35) = 1.76$, $p = .09$. However, this trend was in the opposite direction than predicted by any of the theories; that is, performance was *better* in the hard interruption condition.)

### *3.4.3. Discussion*

The data evidence the Levels-of-Processing effect; memory performance was poorer after the letter-search task than after the imagination or rehearsal task. Better-than-chance performance in the letter-search condition shows that the task resulted in some memory encoding.

More interestingly, the results show no effect of interruption-difficulty. The interrupting task in the hard condition required almost immediate response (2000 ms), which should have disturbed any strategy used to refresh task representations during the interruption task. More interestingly, the absence of the interaction effect between interruption-difficulty and levels-of-processing supports the notion that rehearsal is not naturally utilized as a strategy to compensate for a low level of performance. Activation-based models suggest that low levels of activation should lead to vulnerability to interruptions, unless activation can be built during the interruption lag or the interruption by the means of some strategy, similarly as the tag-based models. The rapid pace in the hard task most probably prevented any kind of simultaneous rehearsal or intermittent building up of activation.

Table 2
Effects of interruption-difficulty in different encoding conditions (Experiment 3)

| Main task encoding condition | Interruption-difficulty effect[a] | | |
|---|---|---|---|
| | Correct recognitions | False recognitions | Omissions |
| Imagine | −.22 | .39 | −.17 |
| Rehearsal | .69 | −.53 | −.17 |
| Letter-pair detect | .00 | .56 | −.56 |

[a] Calculated by subtracting scores in the easy interruption condition from scores in the hard condition.

By contrast, the results are consistent with the LTWM theory. Having to adopt a very weak encoding strategy in the main task does not lead to adoption of rehearsal as the maintenance strategy, because the information is encoded, albeit weakly, in LTM, and can be reinstated with retrieval cues upon task resumption.

## *3.5. Experiment 4*

The first three experiments have pointed out that regardless of their pacing, intensity, or difficulty, interruptions have no negative effect on skilled readers' ability to remember the semantic contents of expository texts. Moreover, rehearsal is not likely to be adopted as means to keep activation high or refresh tags in STWM in this task. The results provide evidence for the main assumption of the LTWM theory, according to which information is safeguarded from interruptions by the means of their encoding to retrieval structures in LTM.

The acid test for the LTWM theory, however, is missing. Namely, a key prediction of the theory is that skilled activities are vulnerable to interruptions in conditions where the time available for encoding, as shaped by the task environment, does *not* match the *speed of encoding to LTWM*. Rapid pace of task processing does not allow for sufficient encoding of retrieval cues and intra-item associations into the retrieval structure, and at the time of task resumption, it is difficult to reinstate this representation to integrate the new incoming material. Another adverse effect might be caused by the currently processed information in fragile state in STWM being practically wiped out by an abrupt interruption demanding an immediate task-switch.

In this experiment, we test this hypothesis by speeding up the main task to approximate the limits of the subjects determined in a pilot experiment. The experimental procedure is otherwise similar to Experiment 1, but presentation time for sentences in the main task has been reduced from 6000 ms (Exp. 1) to only 2500 ms, a practical maximum pace determined in a pilot study.

### *3.5.1. Method*

*3.5.1.1. Participants, materials, design, and procedure.* Twenty-four students from the University of Jyväskylä participated in this experiment in partial fulfillment of course requirements.

All materials, texts, memory test items, and the multiplication task were adopted from Experiment 1.

The procedure was identical to that in Experiment 1, except for the reading pace, which was accelerated from 6000 (Exp. 1) to 2500 ms per sentence. This rate was determined on the grounds of a pilot study ($N = 5$) where we compared 2000, 2500, and 3000 ms. As the participants regarded 2000 ms too difficult and stressing, we decided on using 2500 ms.

In all other respects, the method was identical to Experiment 1.

### *3.5.2. Results*

A single sample *t* test (against 0) for the total memory accuracy score was statistically significant, $t(23) = 9.01$, indicating that the level of performance was above chance level.

There was a significant interruption cost on correct recognitions, $t(23) = -2.17$, $p < .05$, but no effect on false recognitions, $t(23) = .06$. There was a borderline-significant cost of interruption on omissions to the direction predicted by LTWM theory, $t(23) = 1.99$, $p = .06$.

### *3.5.3. Discussion*

Interruption caused a decrease in recognition accuracy, particularly a *decrease* in correct recognitions and a borderline-significant *increase* in omissions, but had no effect on false recognitions. The explanation of the LTWM theory is that because of the fast pace it may be that insufficient time is allocated for encoding to retrieval structures in LTWM, which significantly hampers the ability to reinstate the correct representation upon task resumption to continue integration of new stimulus materials to the task representation.

Interestingly, both activation-based and tag-based models of interruptions also predict an interruption cost for this last experiment (as they do for the three other experiments), although for different reasons. The former holds that less activation can be built because of speeded task processing, which implies increased vulnerability to proactive interference upon task resumption. The finding that interruptions did *not* affect false recognitions is not in line with what activation-based models would have suggested, because decreased activation should have made distinction from pre-existing knowledge in LTM difficult. The latter holds that construction of tags should be compromised by speeded task processing, leading to omissions because of loss of tags. This explanation, therefore, cannot be refuted based on this experiment only, but the evidence brought about the whole set of experiments must be taken into account.

There is, however, one important difference in the predictions between the two alternative models and the LTWM theory. Namely, both alternatives suggest that the main *locus* of the adverse effect of interruption should be in remembering the part of the task processed *before* the interruption. If tags are lost or activation decreased, it should be difficult to distinguish the information stored before the interruption from other information in LTM. Of course, losing pre-interruption information would make integration upon task resumption difficult as well, meaning that a smaller effect should be seen for the part processed after the interruption. It should be smaller because information *after* the interruption can be kept active or tagged as long as no other interruptions emerge that force a task-switch before the recognition test. To sum up, interruption should have an effect either on the pre-interruption part, or on *both* the pre- and the post-interruption parts, but not only on the post-interruption part.

The LTWM theory makes a different prediction. The weak encoding of retrieval structures in speeded processing should lead to two negative effects: first, decreased overall accuracy because of less encoded materials—as was actually witnessed in the poorer overall level of performance of Experiment 4 when compared to Experiment 1; second, interruption-caused impairment mainly in resuming the task when new materials should be integrated to existing ones, and should thus show up as decreased memory accuracy for the post-interruption part. Why does LTWM predict no effect on the part processed *before* the interruption task? Because of the very brief presentation in the main task, encoding resources are already used to the maximum both in interrupted and in non-interrupted conditions, which means that the resulting representations built to the long-term storage are similar in both conditions. (Moreover, there is no lag in the interrupted condition that could be used to safeguard those STWM contents to LTWM just before switching to the interruption task.) To sum up, LTWM predicts a disruptive effect of interruption restricted to the post-interruption part.

We turn to assess this hypothesis in the next subsection.

### 3.5.4. *Post hoc analysis of the locus of interruption effect*

We analysed interruption effects separately for the blocks read before and after the interruption. In line with the predictions of the LTWM theory, there was a significant interruption cost on correct recognitions of items presented *after* the interruption ($M = -1.25$, $\mathrm{SD} = 2.69$), $t(23) = -2.27$, $p < .05$, but *not* on items presented *before* the interruption ($M = -.29$, $\mathrm{SD} = 2.69$), $t(23) = -.53$. The interruption effect on omissions was also restricted to the items presented after the interruption ($M = 1.08$, $\mathrm{SD} = 2.43$), $t(23) = 2.18$, $p < .05$, and there was no significant interruption effect on pre-interruption omissions ($M = .33$, $\mathrm{SD} = 2.66$), $t(23) = .61$. Finally, there were no effects of interruption on the false recognitions before or after the interruption (all $M$s $< .04$), all $|ts(23)| < .15$.

These findings corroborate the predictions of LTWM but not of the two alternative models both of which predicted that there should be a cost on the pre-interruption part.

## 3.6. *Analyses of experiments 1–4*

### 3.6.1. *Power*

Relying on null results in empirical argumentation beckons analysis of experimental power. We wanted to evaluate the probability of observing a negative interruption effect in the three experiments that did not show one. The effect size was determined a posteriori by the effect captured in Experiment 4 where the effect size $f$ for the difference between the conditions for correct recognitions was .44. This effect size is in line with our previous experiment showing an interruption cost in speeded comprehension of expository texts where the effect size was .40 (Oulasvirta and Saariluoma, 2004, Exp. 1). One-way analyses were used, as the theory (as well as the two alternative theories evaluated here) could only explain *decreases* in memory performance due to interruptions. The software used was G*Power (Faul and Erdfelder, 1992). The power analyses for Experiments 1 and 2 yielded a power estimation of .67. Thus, the combined power of these two experiments is

$$1 - (1 - .67)^2 = .89.$$

Calculated analogously, the power of Experiment 3 is .82. The combined power of the three experiments is thus

$$1 - (1 - .89)(1 - .82) = .98.$$

### 3.6.2. *Overall memory performance*

A quick comparison of experiments 1–4 reveals some salient differences in the level of achieved memory accuracy (see Appendix A). Most notably, the proportion of correct recognitions is remarkably lower in Experiment 4 than in Experiments 1 and 2 (about 30 percentage units), although they differed only in respect to presentation speed (Exps. 1 and 2) and materials (Exp. 2). At the first blush, this low rate in Experiment 4 suggests that the obtained interruption effect could be an artefact caused by the level of performance being already close to the floor. However, because of the instructions to avoid false recognitions and the availability of "don't know" response, their performance was actually substantially above the chance level. Had the participants remembered nothing, they could have utilized three strategies: (1) respond only "don't know"; (2) respond randomly "yes" or "no"—leading to correct and false recognitions both being at the .50 level; or (3) respond randomly "yes", "no", or "don't know"—leading to correct and false recognitions and omissions all being at the .33 level. All three strategies would have led to our total memory accuracy measure being 0. Because it was in fact substantially higher, we conclude that the obtained interruption effect is not due to a floor effect.

Secondly, given that performance was significantly above chance level also in Experiment 4, what then explains the differences *between* that and the other experiments? First, we believe that the natural explanation for Experiment 2 having higher level of correct recognitions is that different materials were used (history set). Second, the lower level of correct recognitions in Experiments 3 and 4 is best explained by the experimentally induced limitations to encoding resources in the main task. In Experiment 3, the participants were not reading naturally but were asked to count p-letters, form a mental image, or repeat aloud the sentences, conditions which naturally hamper encoding take away processes (Craik and Lockhart, 1972). Similarly in Experiment 4, as we have explained, presentation speed was so quick that encoding was compromised, also naturally resulting in a lower level of memory.

Finally, when comparing Experiments 1, 3, and 4, which all employed the natural sciences text set (unlike Exp. 2), it can be seen that the effect of limited encoding resources in Experiments 3 and 4 did *not* affect false recognitions but led to an increase in the proportion of omissions from .07 of Experiment 1 to .20 of Experiment 4. As argued, this finding is not in line with temporal cue based accounts which also predict increased proactive interference (here: false recognitions) when less activation can be built for the main task representation. The LTWM theory, by contrast, predicts decreased encoding to retrieval structures, which is reflected as mainly increased omissions in the expense of correct recognitions.

## 4. General discussion

Our starting point in this article was the observation that interruptions are a natural and integral part of everyday multitasking. Letting oneself being interrupted is essential for carrying out multiple tasks effectively in a manner sensitive to both external and internal demands. We seem to have remarkable cognitive machinery that is able to adapt to temporally fragmented task processing so that satisfactory level of performance can be maintained. However, at times this ability is compromised and interruption costs emerge. This paper has addressed a fundamental question distinguishing the two outcomes: how do people distribute information between STM and LTM storages during interrupted task performance?

The first three experiments evidence remarkable robustness of memory in interrupted task processing. This phenomenon, called here interruption tolerance, was demonstrated with two different main task materials. Interrupting reading with a 30-s multiplication-verification task did not hamper memory accuracy even when the difficulty, pacing, or complexity of the interrupting task were high. No costs to memory accuracy were apparent. Only when severe time pressure was imposed to encoding in the main task was a negative interruption effect observed; and even then the negative effect was relatively small (effect size $f$ .44). The locus of this cost was on the part processed *after* the interruption, indicating difficulties in reinstating the main task representation to integrate the incoming information, as hypothesized by the theory of LTWM.

How can these results be explained? We here sketch an elaborated version of how LTWM contributes to the safeguarding of task information. In skilled task processing, people focus on the goals of that task. Initiation of a task creates or re-activates a task representation in LTWM, which contributes to determining the flow of information between memory systems: what is relevant in the presented information, and what intermediate products need to be stored to memory for successful execution of the task. In a task such as reading a book, the representation is essential in guiding eye movement on the page, pacing page turns, encoding and updating the representation of the story contents, and keeping account of which stage of the reading plan we are in. Intermediate products of such processing are incidentally encoded to related retrieval structures that can be retrieved only with certain retrieval cues. It seems plausible that the main function of STWM is to manipulate, transform, and elaborate task-relevant information (Cowan, 2001; Saariluoma, 1995), but not to *store* task information beyond these temporary operations. Thus, no tags are needed to be stored into STWM beyond the task processing, since the task representations can be updated and maintained in LTWM, and activated with proper environmental retrieval cues (see also Trafton et al., 2003). If an interruption takes place, task representation in LTWM remains in an interrupted yet intact state. There can be a disruption of contents in STWM, but most likely people can use the few seconds available upon task-switching to encode some of those contents to LTWM as well. Moreover, in longer tasks such as reading, the proportion of total task information in fragile state in STWM is quite likely small because of rapid encoding to LTWM. Hence, the possible loss of STWM contents due to very abrupt interruption would only pose a small cost to task representations. We could possibly return to the task hours, days, or even weeks later. It is the retrieval cues provided by the environment or generated by the person that activate the relevant task representations from LTWM upon task resumption. Task representations in LTWM are thus "content-addressable" for the time that the LTWM store can be used. LTWM then moves the needed information to STWM whenever active and conscious processing is to take place. When the retrieval cues are not encoded well, or sufficiently, as we argued might happen when the encoding skills do not match the processing demands assumed by the task environment, resumption of the task requires the use of alternative strategies, for example searching for cues in the environment or generating them based on semantic memory on the task. There, the cognitive activity in resumption would be closer to reconstruction than reactivation, a phenomenon we saw in Experiment 4 where the locus of interruption cost was on the part of the task processed after the interruption, not before it.

### 4.1. *Role of rehearsal*

Over the experiments, we have argued that these findings are difficult to account for by the two main competing theoretical models. The first suggests that tasks are kept active by the means of tags in STWM that point to larger chunks of relevant information in LTM (Gobet and Simon, 1996a, b, 1998; Gobet, 1998, 2000a, b; Simon, 1976). The main hypothesis in regard to surviving interruptions is that some sort of rehearsal or refresh strategy is used to keep these tags active in the limited capacity STWM. Within 30 s of interpolated activity without refreshing, the tags should be practically wiped away from STWM, thus making resumption of the task significantly more difficult. If, on the other hand, rehearsal can be somehow exercised during the interruption, fluent resumption is possible because the tags are available.

The four experiments presented here do not support the idea that safeguarding is driven by STWM. Two kinds of evidence support this conclusion. First, memory performance was unaffected by an attention-demanding interruption task (Exps. 1 and 2), even when the attentional demands of the interrupting task were very high (Exps. 2 and 3). Prevention of maintaining tags in STWM should have caused omissions and impaired task resumption. Second, subjects did not engage in strategic maintenance of STWM contents even when they were given a possibility to do it (Exp. 2, self-paced group, and Exp. 3).

There are some reasons to believe that it is unlikely that people would resort to rehearsal as a safeguarding strategy in semantically oriented tasks. Firstly, Naveh-Benjamin and Jonides (1984) suggest that rehearsal consists of two-stages: first the assembling of the items and setting up of an articulatory program, second the repetitive execution of the rehearsal program. Their suggestion is that memory is enhanced mainly at the first stage where reorganization of the materials takes place. Lockhart and Craik (1990) further argued that the second stage, or rote maintenance, of linguistic items is in general associated with an enhancement of the surface (acoustic and articulatory) properties of the materials. This causes problems upon retrieval: "When information has been encoded phonologically or with an unrelated mnemonic, a desired piece of relevant information cannot be accessed by a semantic cue, which forces the subjects to search for the information sequentially" (Ericsson and Delaney, 1999, p. 285). In other words, rehearsal affects cues that might not be easily available upon task resumption. Secondly, the rate of rehearsal is actually quite slow (Baddeley, 1986; Zhang and Simon, 1985), as discussed above. Taken together, then, *the cost-efficiency of rehearsal* as a safeguarding strategy is low. For the effort and time put in rehearsing, only negligible advantages are expected in terms of safeguarding or quicker and more accurate task resumption. For example, Trafton et al. (2003) found that the benefit for 8 s of rehearsal, although voluntarily utilized by the participants, was only 4 s (from 8 to 4 s) quicker task resumption. In line with the LTWM theory, they found that this benefit for rehearsal virtually disappeared with some practice in the task.

Rehearsal should thus be a more prominent strategy in activities where LTM cannot be utilized efficiently, especially in situations where none or only few of the conditions listed in Section 2.2 are fulfilled. For example, in the face of massive interference caused by repeating, highly similar and difficult-to-encode task conditions rehearsal might be a good or even the only available strategy (Altmann, 2002; Brown, 1958; Cowan, 2001; Muter, 1980; Peterson and Peterson, 1959). Moreover, it might have a role when subjects are motivated to perform extremely well and are thus willing to adopt effortful yet poor strategies to enhance performance.

This notion might help to explain a pattern of results in the literature that, initially, seem to contrast our finding that *the processing demands of the interrupting task* do not affect memory for the main task. Gillie and Broadbent (1989) interrupted a computerized task of collecting specified objects from buildings while moving a character in a city. In contrast to our findings, the complexity of the interrupting task affected the interruption cost, which made them to conclude that the amount or *intensity* of information processing of the interrupting task determines disruptiveness. However, their experiment, requiring remembering of arbitrary word lists changing from trial to trial, is a good example of task conditions involving meaningless materials repeated over trials. In these conditions, the use of retrieval structures is unlikely, which quite likely triggered rehearsal as the maintenance strategy. Indeed, Edwards and Gronlund (1998) showed that if the order of the to-be-collected items is not random but fixed over trials, thereby supporting development of retrieval structures, the interruption cost virtually disappears over just six trials.

## 4.2. *Role of temporal cues*

The second alternative model suggests that task information is re-activated upon task resumption from LTM by the means of their higher activation levels or other time-based cues. The main hypothesis is that activation of representation is transient and levels down slowly after processing has been ended (Anderson, 1983; Schneider and Shiffrin, 1977). After a long interruption, activation is at so low level that proactive interference should occur. This means that there are practically two ways of surviving the interruption: (1) by building enough activation already in the main task or (2) refreshing the activation during the interruption or the interruption lag. As we have argued, the second option (active maintenance) is a quite implausible explanation for our data, because our interruption task was sufficiently demanding to have prevented any intermittent rehearsal. Moreover, subjects switched to the interrupting task so quickly that practically no efficient retrospective rehearsal could have been carried out (Exp. 2).

Our experiments show no evidence for the first option either—the hypothesis that temporal cues (such as activation) would have been utilized to retrieve the task representations upon task resumption. That no costs were observed in two experiments using different materials and an interruption as long as 30 s is difficult to explain by this hypothesis. However, the psychological reality of decay of memory traces or temporal cues is not denied in the LTWM theory, it is only claimed that because of their weakness as retrieval cues their use as the primary cue is avoided if possible. "In those cases where temporal information is used to distinguish the most recently associated piece of information to a retrieval cue, an interruption will increase the effects of proactive interference and this might disable a successful reinstatement of all relevant information" (Ericsson and Delaney, 1999, p. 268). Altmann (2001) has proposed an extended model of programmer's LTWM that utilizes episodic indexing of events as one mechanism of efficient re-generation.

The question then remains in what situations people would, in general, use temporal instead of semantics-driven maintenance. There are at least two important uses for temporal cues known well in memory psychology. Recency and primacy effects seem to be pervasive memory phenomena that also appear in expert memory (Baddeley, 1986; Ericsson and Kintsch, 1995). People can learn to use such recency and primacy cues also to *inhibit and forget* irrelevant task information in repetitive task conditions where only the first or last presented information matters, like for example when observing speed limit information while driving (Altmann, 2002). Thus, from the LTWM perspective, this reliance on temporal cues might be induced by the demands of the task to be selective.

The question is intertwined with the idea that the *length* of the interruption is associated with its disruptiveness. We maintain that length should matter only if temporal cues are used in retrieval. For example, in the task of Gillie and Broadbent (1989) that we argued shows use of STWM-based rehearsal as the maintenance strategy, manipulating the length did *not* affect task resumption latency. However, in conditions of massive interference *and* where rehearsal is ruled out by experimental manipulation, length of the interrupting period is indeed shown to matter. In classic studies with wordlist remembering, dramatic forgetting of the main task has been observed even after few seconds when shifting attention to another task (e.g., Brown, 1958; Glanzer et al., 1984; Keppel and Underwood, 1962; Muter, 1980; Peterson and Peterson, 1959).

To summarize, interruption tolerance is possible in some restricted yet practically important conditions (reviewed in Section 2) where retrieval structures of LTM can be efficiently utilized to avoid the adverse effects of interruptions, but this does not mean that LTWM could be used in all conditions. In other words, LTWM best captures skilled performance in some everyday tasks, and rehearsal and temporal cues both might have important role as interruption strategies in other tasks.

### 4.3. *Qualities of task environments that support interruption tolerance*

The question of interruptions has become acutely important in the era of modern human–computer interaction (McFarlane and Latorella, 2002). Computer-triggered events at the PC desktop, such as dialogues and pop-ups, screen savers, advertisements and banners, instant messages and intelligent agents, notifications and reminders, multimedia, mobile and ordinary phones, and colleagues can all divert attention to a message that is irrelevant from the point of view of the interrupted main task. Modern PCs have worsened this situation by providing a wider selection of applications, richer information displays, and operating systems that support fluent simultaneous execution of programs (Card and Henderson, 1987). The situation is similar, or even worse, in ubiquitous computing, where external displays, wearable computers, physical computing artifacts, and mobile devices compete for users' attention (Intille, 2002; Oulasvirta and Salovaara, 2004; Vertegaal, 2003; Weiser, 1991, 1993). In mobile interaction, the processing of an interactive task is shown to fragment into bursts of just few seconds per turn (Oulasvirta et al., 2005).

In this paper we have applied the LTWM theory to investigate interruptions from a *user psychological* point of view (Oulasvirta and Saariluoma, 2004; Saariluoma, 2005). Since interruptions are very typical in normal human–machine interaction (e.g., Mark et al., 2005; O'Conaill and Frohlich, 1995), these kinds of user psychological investigations can be made, more effectively and in a more argumentative manner, to utilize psychological knowledge in design and engineering. It seems that in addition to solving some discrepancies in the body of interruption research, the theory of LTWM provides good possibilities to understand users' memory processes during interruptions even to the level where making suggestions for task design is possible.

Along these lines, we turn to look at three general principles of task environment design that can be used to support interruption tolerance. Generally speaking, the contribution of the theory of LTWM to the design of non-disruptive interfaces comes from the suggestions of how interfaces could support robust encoding and efficient re-activation of the main task. This approach complements the prevailing approaches that have focused mainly on the interrupting task, its content and timing.

#### 4.3.1. *Support for development of skilful encoding*

Our main suggestion is to look at *main task's support for memory skills* that enable users to overcome interruptions without adverse effects, instead of putting effort to scheduling, timing, or content design of interruption tasks that has been the prevailing approach (cf. Field, 1987; Kreifeldt and McCarty, 1981). As suggested by Edwards and Gronlund (1998), "a well-designed task environment might allow for the creation of associative connections among task components that would result in a mental representation for the task that was relatively immune to the adverse effects of interruptions."

What qualities of task environments help the development of LTWM skills? Drawing from the conditions of LTWM encoding reviewed in Section 2, three main qualities can be recognized:

(1) *Consistency and constraints*: Consistency in how materials are presented to the user is crucial. Practice in consistent and constrained circumstances supports the adaptation of memory skills to retrieval demands posed by the interface (see also Vicente and Wang, 1998). The importance of consistency in supporting interruption tolerance was well illustrated in an experiment of

Edwards and Gronlund (1998) evidencing elimination of interruption cost due to consistency in the presentation of the main task materials.

(2) *Predictability*: By "predictability," we mean that the interface makes it apparent what is going to take place in the near future. In temporally ordered procedural tasks such as installation programs on PCs, users benefit from seeing what is to be expected next. Also in more dynamic application domains, if future events can be deduced or guessed, visualizing them for the user helps both to (1) encode for anticipated future retrieval demands (e.g., encoding information that is needed during the next step of the procedure) and (2) pace and control the encoding according to temporal or other restrictions (e.g., finish writing a sentence before switching to an email application to reply to an urgent email).

(3) *Transparency*: "Transparency" refers to visualizing the working mechanism and current state of the application to the user. Similarly to predictability, transparency supports one anticipating future events and controlling the encoding operations accordingly.

### 4.3.2. *Match to encoding speed*

The issue of available vs. needed time is crucial, because many aspects of time can and are influenced by the properties of the task environment (Hollnagel and Woods, 2005). Rhythms, paces, turns, schedules, division of tasks to subtasks and phases, and timing of events all affect the availability of time during task processing.

*Encoding speed* denotes the time required for transferring processed information to LTWM. Several studies have shown the increase of encoding speed due to development of memory skills with practice (e.g., Chase and Ericsson, 1982; Ericsson and Kintsch, 1995; Ericsson and Polson, 1988; Ericsson and Staszewski, 1989). The key suggestion here is matching the task processing demands to the encoding speed of the user. If the required speed is slower than the processing speed demanded by the task, an interruption cost can emerge, as shown in Experiment 4. Moreover, extreme time pressure in the task may lead to selection of stimulus-specific processes at the expense of processes relying on a more general representation (Ericsson and Delaney, 1999). The suggestion is in line with the general human factors principle of designing tasks to match users' capabilities and skills.

Before going into concrete implications to interface design, we turn to consider the implications of the notion of encoding speed to *timing of interruptions* in general. Namely, one factor that has been studied recently is the timing of interruptions in regard to the *phase/stage* of the main task. Many of the previous studies have concentrated on comparing different mental or task stages during which interruption occurs (Adamczyk and Bailey, 2004; Ho and Intille, 2005; Iqbal et al., 2005; Monk et al., 2004). For example, stimulated by the hypothesis put forward in 1986 by Miyata and Norman (1986), Cutrell et al. (2000) found that interruptions to web search were most disruptive when they occurred during the evaluation of search results, but not when planning the search or executing it (see also Monk et al., 2004). They argued that evaluation requires re-activating the position in the search results list upon task resumption, whereas the two other phases are less dependent on memory. Consequently, some researchers have researched the possibility that it is *memory load* instead of the phase per se that affects disruptiveness. Shifting from surface features of the task to mentalistic explanations have unfortunately failed to provide uniform evidence in this case. In Edwards and Gronlund's (1998) experiment's low memory condition, only six locations had to be searched in the Gillie-Broadbent task, in comparison to 19 in the high memory condition. To their surprise, they found no difference between the high and low memory load conditions. By contrast, Detweiler et al. (1994), found interruption cost to be most pronounced with larger working memory loads. Here, results from different experimental paradigms exhibit a clear discrepancy that the LTWM theory can attempt to explain.

From the perspective of the LTWM theory, memory load or task phase are *not* the factors determining disruptiveness to interruptions although they do coincide with it. Chess experts, for example, in playing multiple games of chess blindfolded, can play and later remember more games more accurately than novices, indicating *larger* memory load—yet they are *less* disrupted by interruptions (Charness, 1976; Saariluoma, 1991). Instead of memory load, as measured by units of information processed in memory in a unit of time, we posit that the key factor is how quickly that particular information can be safeguarded. Each task phase and materials is associated with different encoding goals, available resources, pre-existing knowledge, and encoding speed. For example, some scientists are more familiar with Method sections of articles and can more easily grasp their ideas and tie them into their existing retrieval structures than those of Conclusions, which are always unique to the paper. Unfamiliarity of those contents would mean that during the period of reading them, encoding speed is lower and processing thus more vulnerable to interruptions. These within-task differences can explain why interruptions "in the middle" of some tasks have more negative effects than in some other tasks, and why interruptions between tasks are less detrimental than during them. Unless demands and skills match, vulnerability to interruptions can occur as we saw in Experiment 4.

Whereas determining available time might be straightforward from an interface, the question then arises how to determine LTWM encoding speed, a latent mental factor? Encoding speed, as measured in terms of units of encoded material per a time unit, is dependent on the processing skill and resources of the subject interacting with properties of the material to be remembered. For this reason, speed can vary over an enormous range. Some upper boundaries for encoding speed can be, however, estimated from task performance. For example, storing chess positions in memory entails encoding of the whole chess board, including the position of all 24 pieces, in time less than 5 s, which yields an encoding time of about 200 ms per piece (Ericsson and Delaney, 1999). In a similar fashion Ericsson and Polson (1988) estimated that JC, an expert waiter, encoded menu items with the speed of about 1.5–2.0 s per item.

In addition to matching estimated encoding speed to available time in the user interface, and supporting the development of rapid and skilful encoding as suggested in the previous subsection, several general measures can be taken to improve the way *time can be used as a resource* in task encoding:

(1) *Short interaction chains*: By keeping interaction chains leading to a goal or subgoal short and concise, a designer can increase the probability that a user's encoding of a chunk of a task has taken place as a whole before the interruption occurs, thus decreasing the possibility of materials being in a fragile state in STWM. For example, in mobile applications, where interruptions are typically more common and harmful than in desktop applications, one rule of thumb among practitioners has been to design interaction chains that last less than 20 seconds (Brown, 2004, personal communication 2004).

The last four relate to interactive aspects of user interfaces:

(2) *Interruption lag*: Providing an interruption lag means that the task privileges some time just before switching to the interrupting task for finishing the encoding of task-relevant information to LTWM. Interpreted simplistically, implementing an interruption lag would mean imposing a lag as some kind of "freezing" of the UI, which in many cases would be unnatural and annoying. As argued above, the cost-efficiency of interruption lags is quite low. A more sophisticated option is enabling user-control over the timing of task-switches (see below).

(3) *Interruption locks*: By "interruption locks" we mean designing a switching strategy that protects certain "memory critical" operations by disallowing intermittent interruptions (Adamczyk and Bailey, 2004; Iqbal et al., 2005). These types of practices have been adopted for cockpits (e.g., for certain checklists) and control rooms and they could be integrated as well into more traditional desktop UIs as well. What would be locked are those tasks or subtasks that involve much slower encoding speed than what the interaction allows for.

(4) *User-pacing of interruptions*: User-pacing here means that the user can decide when (and possibly how) the interruption takes place. McFarlane (2002) has proposed two forms of user-control over the timing of interruptions: scheduled and negotiated. In the scheduled condition, interruptions occur within predictable time intervals (e.g., every 25 s), which makes preparing for them easier also from the encoding point of view. In the negotiated condition, a signal on the user interface marks the appearance of the interrupting task, but the user can decide when to switch to it (self-pacing). The Scope interface representing incoming communications (e.g., emails and instant messaging) by van Dantzich et al. (2002) provides an example in which incoming messages are visualized according to their priority and recency on a radar-like display. Solutions like these provide the user with the possibility to switch to the interrupting task when encoding is finished in the main task. However, more work is needed to understand the strategies users typically adopt to monitor and manage interruptions in such situations. It might prove that they are not optimal but opportunistic, and thus maybe harmful to the encoding of the main task (e.g., Eyrolle and Cellier, 2000; O'Conaill and Frohlich, 1995). An interesting research question for future work arises: how can interfaces promote the adoption of switching strategies that match encoding speed for the task.

(5) *User-pacing of main task*: User pacing of main task means that the user can control the durations of and shifts between task phases. If various pacing and switch strategies can be supported, users should be able to match their encoding skills to presentation time and timing of phase switches.

### 4.3.3. *Encoding-retrieval symmetry*

Our third and final class of implications concerns the availability of retrieval cues upon task resumption. There, information in LTWM can be retrieved only with the relevant cues (see also Trafton et al., 2003), and not all the cues available during some cognitive processing are also accessible upon task resumption (see also Morris et al., 1977; Tulving and Thomson, 1973). The challenge for design, then, is to ensure that those features of the task materials or environment encoded are also available and distinctive enough at that time.

We divide our suggestions to two:

(1) *Support semantic encoding and access*: By "semantic access," we mean prioritizing semantic organization of user interface. Similar suggestions have been made recently by Altmann (2001) and Terrier and Cellier (1999), but the concept has remained somewhat vague. By semantic organization of interface we mean that materials are organized according to some semantically meaningful principle in the interface. For example, providing short summaries of texts in reading-oriented applications not only supports the user's understanding the materials better and thus developing the encoding skills related to the material itself, but also helps in task resumption by working as quickly accessible retrieval cues. In addition, apparently non-textual materials can also benefit from semantic organization. For example, links need not be arbitrarily positioned on a panel on a

web page: the user can benefit from a higher-level organization in which links are organized semantically in groups and the groups are positioned in relation to each other according to a higher-level concept that is made visible to and understood by the user (e.g., www.cnn.com/world organizes world news according to region). Semantic access is in contrast to "surface access" which means that materials on an interface are organized according to some perceptual or surface feature. In some cases, this organization leads to a requirement for the user to encode positions of materials, which we have seen can easily lead to interruption costs caused by slower encoding speed. For example, a visual marker of a to-be-evaluated database search engine results page was found of no use in a study by Cutrell et al. (2001). (However, Cutrell et al.'s aim for ecological validity may have resulted in little experimental power.) Textual cues, on the other hand, have proved useful in the studies of semantically oriented main tasks like reading (Glanzer et al., 1984).

(2) *Interactive reactivation*: By "interactive reactivation" we mean providing interactive support for the user to find those particular cues that help reactivating the correct mental representation after an interruption. For example, "focus + context" visualizations in data-driven applications or navigation histories in navigation-oriented applications help users not only understand the domain better but also provide interactive support for developing strategies for finding cues that help in reactivating the correct task representations.

## Acknowledgements

We thank Pauli Salo, Jukka Purma, Jarmo Herkman, and Anu Pesonen for co-operation in recruiting participants from their courses, and Virpi Kalakoski, Sebastian Thärman, Juhani E. Lehto, Greg Trafton, and two anonymous reviewers for their valuable comments. We are also indebted to Petri Ojala, Pekka Toppi, and Rami Honkanen, who helped running the experiments. This work has been financially supported by the Academy of Finland, TeliaSonera, Nokia, TKK:n Tukisäätiö, HPY:n Tutkimussäätiö, Alma Media, Elisa Communications, Swelcom, and TEKES.

## Appendix A

Memory accuracy data from Experiments 1, 2, 3, and 4 is shown in Table A1.

Table A1

| Experiment | Interruption condition | Response measure* | | |
|---|---|---|---|---|
| | | Correct recognitions | False recognitions | Omissions |
| 1 | Interrupted | 23.1 (64%) | 10.2 (28%) | 2.7 (8%) |
| | Continuous | 23.6 (66%) | 10.2 (28%) | 2.2 (6%) |
| 2 | Self-paced group | | | |
| | Interrupted | 27.2 (76%) | 5.0 (14%) | 3.8 (10%) |
| | Continuous | 27.0 (75%) | 5.6 (16%) | 3.4 (9%) |
| | Experimenter-paced group | | | |
| | Interrupted | 29.4 (81%) | 4.1 (12%) | 2.5 (7%) |
| | Continuous | 27.8 (77%) | 5.5 (15%) | 2.8 (8%) |
| 3 | Easy | 18.3 (51%) | 9.6 (26%) | 8.2 (23%) |
| | Hard | 18.7 (52%) | 9.9 (28%) | 7.3 (20%) |
| 4 | Interrupted | 17.5 (48%) | 10.9 (30%) | 7.6 (22%) |
| | Continuous | 19.0 (53%) | 10.8 (30%) | 6.1 (17%) |

Figures represent mean number of responses (max. 36, i.e. 12 test items × 3 texts) in that condition for a subject. Percentages of all responses are given in parentheses.

## References

Adamczyk, P.D., Bailey, B.P., 2004. If not now, when? The effects of interruption at different moments within task execution. Human Factors in Computing Systems: Proceedings of CHI'04. ACM Press, New York, pp. 271–278.

Akin, Ö., 1986. The Psychology of Architectural Design. Pion, London, UK.

Altmann, E.M., 2001. Near-term memory in programming: A simulation-based analysis. International Journal of Human-Computer Studies 54, 189–210.

Altmann, E.M., 2002. Functional decay of memory for tasks. Psychological Research 66, 287–297.

Anderson, J.R., 1983. The Architecture of Cognition. Harvard University Press, Cambridge, MA.

Baddeley, A., 1986. Working Memory. Cambridge University Press, Cambridge, UK.

Baddeley, A., 2000. The episodic buffer: a new component of working memory? Trends in Cognitive Sciences 4 (11), 417–423.

Baddeley, A., Chincotta, D., Adlam, A., 2001. Working memory and the control of action: evidence from task switching. Journal of Experimental Psychology: General 130 (4), 641–657.

Bradshaw, G.L., Anderson, J.R., 1982. Elaborative encoding as an explanation of levels of processing. Journal of Verbal Learning and Verbal Behavior 21, 165–174.

Braun, K.A., Rhiannon, E., Loftus, E.F., 2002. Make my memory: how advertising can change our memories of the past. Psychology of Marketing 19 (1), 1–23.

Brown, J., 1958. Some tests of the decay theory of immediate memory. Quarterly Journal of Experimental Psychology 10, 12–21.

Brown, T., 2004. The future of designing experiences. Closing Plenary Speech at Human Factors in Computing Systems: ACM CHI'04, Vienna, Austria, April 24–29.

Byrne, M.D., Anderson, J.R., 1999. Serial modules in parallel: the psychological refractory period and perfect time-sharing. Psychological Review 108 (4), 847–869.

Card, S.K., Henderson, A., 1987. A multiple, virtual-workspace interface to support user task switching. ACM SIGCHI Bulletin 17, 53–59.

Charness, N., 1976. Memory for chess positions: resistance to interference. Journal of Experimental Psychology: Human Learning and Memory 2, 641–653.

Charness, N., 1979. Components of skill in bridge. Canadian Journal of Psychology 33, 1–16.

Charness, N., 1988. Expertise in chess, music, and physics: A cognitive perspective. In: Obler, L.K., Fein, D.A. (Eds.), The Exceptional Brain: Neuropsychology of Talent and Special Abilities. Guilford Press, New York, pp. 399–426.

Charness, N., 1989. Expertise in chess and bridge. In: Klahr, D., Kotovsky, K. (Eds.), Complex Information Processing: The Impact of Herbert A. Simon. Erlbaum, Hillsdale, NJ, pp. 183–208.

Charness, N., 1992. The impact of chess research on cognitive science. Psychological Research 54, 4–9.

Chase, W.G., Ericsson, K.A., 1982. Skill and working memory. In: Bower, G.H. (Ed.), The Psychology of Learning and Motivation, vol. 16. Academic Press, New York, pp. 1–58.

Chase, W.G., Simon, H.A., 1973. The mind's eye in chess. In: Chase, W.G. (Ed.), Visual Information Processing. Academic Press, New York, pp. 215–281.

Cowan, N., 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral & Brain Sciences (Special Issue) 24 (1), 87–185.

Craik, F.I.M., Lockhart, R.S., 1972. Levels of processing: a framework for memory research. Journal of Verbal Learning and Verbal Behaviour 11, 671–684.

Cutrell, E., Czerwinski, M., Horvitz, E., 2000. Effects of instant messaging interruptions on computing tasks. Extended Abstracts of CHI '2000, Human Factors in Computing Systems. ACM press, The Hague, pp. 99–100.

Cutrell, E., Czerwinski, M., Horvitz, E., 2001. Notification, disruption and memory: effects of messaging interruptions on memory and performance. In: Hirose, M. (Ed.), Human-Computer Interaction INTERACT '01. IOS Press (for IFIP), Tokyo, pp. 263–269.

van Dantzich, M., Robbins, D., Horvitz, E., Czerwinski, M., 2002. Scope: providing awareness of multiple notifications at a glance. Proceedings of Advanced Visual Interfaces 2002.

Detweiler, M.C., Hess, S.M., Phelps, M.P., 1994. Interruptions and working memory. Unpublished Technical Report. The Pennsylvania State University, University Park, PA.

Djakov, I.N., Petrovsky, N.B., Rudik, P.A., 1926. Psihologia Shahmatnoi Igry (Chess Psychology). Avtorov, Moskow.

Ebbinghaus, H., 1885. On memory. Dover, New York (Re-printed in 1964).

Edwards, M.B., Gronlund, S.D., 1998. Task interruption and its effects on memory. Memory 6 (6), 665–687.

Egan, D.E., Schwartz, B.J., 1979. Chunking in recall of symbolic drawings. Memory & Cognition 4, 541–547.

Emerson, M.J., Miyake, A., 2003. The role of inner speech in task switching: a dual-task investigation. Journal of Memory and Language 48, 148–168.

Engle, R.W., Bukstel, L., 1978. Memory processes among bridge players of different expertise. American Journal of Psychology 91, 673–689.

Ericsson, K.A., Charness, N., 1994. Expert performance: its structure and acquisition. American Psychologist 49, 725–747.

Ericsson, K.A., Delaney, P.F., 1999. Long-term working memory as an alternative to capacity models of working memory in everyday skilled performance. In: Miyake, A., Shah, P. (Eds.), Models of Working Memory: Mechanisms of Active Maintenance and Executive Control. Cambridge University Press, Cambridge, UK, pp. 257–297.

Ericsson, K.A., Kintsch, W., 1995. Long-term working memory. Psychological Review 102, 211–245.

Ericsson, K.A., Kintsch, W., 2000. Shortcomings of generic retrieval structures with slots of the type that Gobet (1993) proposed and modeled. British Journal of Psychology 91, 571–588.

Ericsson, K.A., Lehmann, A.C., 1996. Expert and exceptional performance: Evidence for maximal adaptation to task constraints. Annual Review of Psychology 47, 273–305.

Ericsson, K.A., Polson, P., 1988. Experimental analysis of the mechanisms of memory skill. Journal of Experimental Psychology: Learning, Memory and Cognition 14, 305–316.

Ericsson, K.A., Staszewski, J.J., 1989. Skilled memory and expertise: mechanisms of exceptional performance. In: Klahr, D., Kotovsky, K. (Eds.), Complex Information Processing: The Impact of Herbert A. Simon. Lawrence Erlbaum, Hillsdale, NJ, pp. 235–267.

Eyrolle, H., Cellier, J.-M., 2000. The effects of interruptions in work activity: field and laboratory results. Applied Ergonomics 31 (5), 537–543.

Faul, F., Erdfelder, E., 1992. GPOWER: A Priori, post-hoc, and Compromise Power Analyses for MS-DOS [Computer program]. Bonn University, Department of Psychology, Bonn, FRG.

Field, G.E., 1987. Experimentus interruptus. ACM SIGCHI Bulletin 19 (2), 42–46.

Fischer, B., Glanzer, M., 1986. Short-term storage and the processing of cohesion during reading. The Quarterly Journal of Experimental Psychology 38 (A), 431–460.

Frey, P.W., Adesman, P., 1976. Recall memory for visually presented chess positions. Memory & Cognition 4, 541–547.

Gardiner, J.M., Richardson-Klavehn, A., 2000. Remembering and knowing. In: Tulving, E., Craik, F.I.M. (Eds.), Handbook of Memory. Oxford University Press, Oxford.

Gilhooly, K.J., Wood, M., Kinnear, P., Green, C., 1988. Skill in map reading and memory for maps. Quarterly Journal of Experimental Psychology 40A, 87–107.

Gillie, T., Broadbent, D., 1989. What makes interruptions disruptive? A study of length, similarity, and complexity. Psychological Research 50, 243–250.

Glanzer, M., Dorfman, D., Kaplan, B., 1981. Short-term processing in the processing of text. Journal of Verbal Learning and Verbal Behavior 20, 656–670.

Glanzer, M., Fischer, B., Dorfman, D., 1984. Short-term storage in reading. Journal of Verbal Learning and Verbal Behavior 23, 467–486.

Gobet, F., 1998. Expert memory: a comparison of four theories. Cognition 66 (2), 115–152.

Gobet, F., 2000a. Some shortcomings of long-term working memory. British Journal of Psychology 91, 551–570.

Gobet, F., 2000b. Retrieval structures and schemata: a brief reply to Ericsson and Kintsch. British Journal of Psychology 91, 591–594.

Gobet, F., 2001. Chunks hierarchies and retrieval structures. Comments on Saariluoma and Laine. Scandinavian Journal of Psychology 42, 149–155.

Gobet, F., Simon, H.A., 1996a. Templates in chess memory: a mechanism for recalling several boards. Cognitive Psychology 31 (1), 1–40.

Gobet, F., Simon, H.A., 1996b. Recall of rapidly presented random chess positions is a function of skill. Psychonomic Bulletin and Review 3, 291–313.

Gobet, F., Simon, H.A., 1998. Expert chess memory: revisiting the chunking hypothesis. Memory 6, 225–255.

Graesser, A.C., Millis, K.K., Zwaan, R.A., 1997. Discourse comprehension. Annual Review of Psychology 48, 163–189.

Green, G.A., Gilhooly, J.K., 1992. Empirical advances in expertise research. In: Keane, M.T., Gilhooly, J.K. (Eds.), Advances in Psychology of Thinking. Harvester Wheatsheat, London, UK.

de Groot, A.D., 1965. Thought and Choice in Chess. Mouton, The Hague.

de Groot, A.D., 1966. Perception and memory vs. thought: some old questions and recent findings. In: Kleinmuntz, B. (Ed.), Problem Solving. Wiley, New York.

Hatta, T., Hirose, T., Ikeda, K., Fukuhara, H., 1989. Digit memory of Soroban experts: evidence of utilization of mental imagery. Applied Cognitive Psychology 3, 23–33.

Ho, J., Intille, S.S., 2005. Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In: Human Factors in Computing Systems: Proceedings of CHI'05. ACM Press, New York, NY, pp. 909–918.

Hollnagel, E., Woods, D.D., 2005. Joint Cognitive Systems: Foundations of Cognitive Systems Engineering. CRC Press/Taylor & Francis, Boca Raton, FL.

Hyde, T.S., Jenkins, J.J., 1969. Differential effects of incidental tasks on the organization of recall of a list of highly associated words. Journal of Experimental Psychology 82, 472–481.

Intille, S.S., 2002. Change blind information display for ubiquitous environments. Proceedings of the Fourth International Conference Ubiquitous Computing (Lecture Notes in Computer Science, vol. 2498), pp. 91–106.

Iqbal, S.T., Adamczyk, P.D., Zheng, X.S., Bailey, B.P., 2005. Towards an index of opportunity: Understanding changes in mental workload during task execution. In: Human Factors in Computing Systems: Proceedings of CHI'05. ACM Press, New York, pp. 311–320.

Kalakoski, V., Saariluoma, P., 2001. Taxi drivers' exceptional memory of street names. Memory & Cognition 29 (4), 634–638.

Keppel, G., Underwood, B.J., 1962. Proactive inhibition in short-term retention of single items. Journal of Verbal Learning and Verbal Behavior 1, 153–161.

Kinnear, P., Wood, M., 1987. Memory for topographic contour maps. British Journal of Psychology 78, 395–402.

Kintsch, W., 1988. The role of knowledge in discourse comprehension: a construction-integration model. Psychological Review 95, 163–182.

Kreifeldt, J.G., McCarty, M.E., 1981. Interruption as a test of the user-computer interface. Proceedings of the 17th Annual Conference on Manual Control. California Institute of Technology, Jet Propulsion Laboratory Publication 81–95, pp. 655–667.

Lane, D.M., Robertson, L., 1979. The generality of levels of processing hypothesis: An application to memory for chess positions. Memory & Cognition 7, 253–256.

Lockhart, R.S., Craik, F.I.M., 1990. Levels of processing: a retrospective commentary on a framework for memory research. Canadian Journal of Psychology 44 (1), 87–112.

Mark, G., González, V. M., Harris, J., 2005. No task left behind? Examining the nature of fragmented work. Proceedings of CHI'05. ACM Press, New York, pp. 321–330.

McFarlane, D.C., 2002. Comparison of four primary methods for coordinating the interruption of people in human–computer interaction. Human-Computer Interaction 17 (1), 63–139.

McFarlane, D.C., Latorella, K.A., 2002. The scope and importance of human interruption in human–computer interaction design. Human-Computer Interaction 17 (1), 1–61.

McKeithen, K.B., Reitman, J.S., Rueter, H.H., Hirtle, S., 1981. Knowledge organization and skill differences in computer programmers. Cognitive Psychology 13, 307–325.

Miyata, Y., Norman, D., 1986. Psychological issues in support of multiple activities. In: Norman, D.A., Draper, S.W. (Eds.), User Centered Systems Design: New Perspectives on Human-Computer Interaction. Lawrence Erlbaum Associates, Hillsdale, pp. 265–284.

Monk, C.A., Boehm-Davis, D.A., Trafton, J.G., 2004. Recovering from interruptions: implications for driver distraction research. Human Factors 46 (4), 650–663.

Morris, C.D., Bransford, J.D., Franks, J.J., 1977. Levels of processing versus transfer appropriate processing. Journal of Verbal Learning and Verbal Behaviour 16, 519–533.

Muter, P., 1980. Very rapid forgetting. Memory & Cognition 8, 174–179.

Mylers-Worsley, M., Johnston, W.A., Simons, M.A., 1988. The influence of expertise on X-ray image processing. Journal of Experimental Psychology: Learning, Memory and Cognition 14, 553–557.

Naveh-Benjamin, M., Jonides, J., 1984. Maintenance rehearsal: a two-component analysis. Journal of Experimental Psychology: Learning, Memory, and Cognition 10, 369–385.

O'Conaill, B., Frohlich, D., 1995. Timespace in the workplace: dealing with interruptions. Proceedings of CHI'95, ACM Press, New York, pp. 262–263.

Oulasvirta, A., 2004. Task-processing demands and memory in web interaction: a levels-of-processing approach. Interacting with Computers 16 (2), 217–241.

Oulasvirta, A., Saariluoma, P., 2004. Long-term working memory and interrupting messages in human–computer interaction. Behaviour and Information Technology 23 (1), 53–64.

Oulasvirta, A., Salovaara, A., 2004. A cognitive meta-analysis of design approaches to interruptions in intelligent environments. Proceedings of ACM Conference on Human Factors in Computing Systems (CHI'04): Extended Abstracts, ACM Press, New York, pp. 1155–1158.

Oulasvirta, A., Tamminen, S., Roto, V., Kuorelahti, J., 2005. Interaction in 4-second bursts: The fragmented nature of attentional resources in mobile HCI. Proceedings of ACM Conference on Human Factors in Computing Systems (CHI'05), ACM Press, New York, pp. 919–928.

Peterson, L.R., Peterson, M.J., 1959. Short-term retention of individual verbal items. Journal of Experimental Psychology 58, 193–198.

Reitman, J.S., 1976. Skilled perception in go: deducing memory structures from inter-response times. Cognitive Psychology 8, 336–356.

Richman, H.B., Staszewski, J.J., Simon, H.A., 1995. Simulation of expert memory using EPAM IV. Psychological Review 102 (2), 305–330.

Royer, J.M., Greene, B.A., Sinatra, G.M., 1987. The Sentence Verification Technique: a practical procedure that teachers can use to develop their own reading and listening comprehension tests. Journal of Reading 30 (5), 414–422.

Saariluoma, P., 1985. Chess players' intake of task relevant cues. Memory & Cognition 13, 385–391.

Saariluoma, P., 1989. Chess players' recall of auditorily presented chess positions. European Journal of Cognitive Psychology 1, 309–320.

Saariluoma, P., 1991. Aspects of skilled imagery in blindfold chess. Acta Psychologica 77 (1), 65–89.

Saariluoma, P., 1995. Chess Players' Thinking: A Cognitive Psychological Approach. Routledge, London, UK.

Saariluoma, P., 2005. Explanatory frameworks in interaction design. In: Pirhonen, A., Isomäki, H., Roast, C., Saariluoma, P. (Eds.), Future Interaction Design. Springer, London.

Saariluoma, P., Kalakoski, V., 1997. Skilled imagery and long-term working memory. American Journal of Psychology 110, 117–201.

Saariluoma, P., Kalakoski, V., 1998. Apperception and imagery in blindfold chess. Memory 6, 67–90.

Schneider, W., Shiffrin, R.M., 1977. Controlled and automatic human information processing: 1 Detection, Search, and Attention. Psychological Review 84, 1–66.

Shneiderman, B., 1976. Exploratory experiments in programmer behavior. International Journal of Computer and Information Sciences 5, 123–143.

Simon, H.A., 1976. The information storage system called human memory. In: Rosenzweig, M., Bennett, E. (Eds.), Neural Mechanisms of Learning and Memory. MIT Press, Cambridge, MA.

Sloboda, J., 1976. Visual perception of musical notation: registering pitch symbols in memory. Quarterly Journal of Experimental Psychology 28, 1–16.

Terrier, P., Cellier, J.-M., 1999. Depth of processing and design-assessment of ecological interfaces: task analysis. International Journal of Human-Computer Studies 50, 287–307.

Trafton, J.G., Altmann, E.M., Brock, D.P., Mintz, F.E., 2003. Preparing to resume an interrupted task: effects of prospective goal encoding and retrospective rehearsal. International Journal of Human-Computer Studies 58, 583–603.

Tulving, E., Thomson, D.M., 1973. Encoding specificity and retrieval processes in episodic memory. Psychological Review 80, 352–373.

Vertegaal, R., 2003. Attentive User Interfaces. Communications of the ACM 46 (3), 30–33.

Vicente, K.J., 1988. Adapting the memory recall paradigm to evaluate interfaces. Acta Psychologica 69, 249–278.

Vicente, K.J., de Groot, A.D., 1990. The memory recall paradigm: straightening out the historical record. American Psychologist 45, 285–287.

Vicente, K.J., Wang, J.H., 1998. An ecological theory of expertise effects in memory recall. Psychological Review 105, 33–57.

Weiser, M., 1991. The computer for the 21st century. Scientific American 265 (3), 66–75.

Weiser, M., 1993. Some computer science issues in ubiquitous computing. Communications of the ACM 36 (7), 75–84.

Zhang, G., Simon, H.A., 1985. STM capacity for Chinese words and idioms: Chunking and acoustical loop hypotheses. Memory & Cognition 13, 193–201.