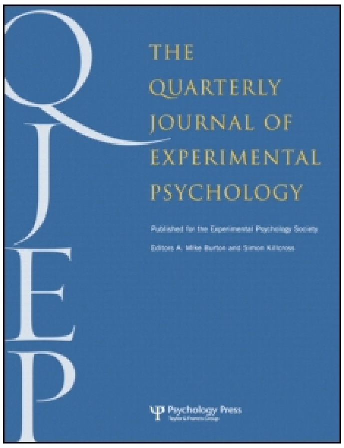
This article was downloaded by: *[Turku University Library]*
On: *27 January 2009*
Access details: *Access Details: [subscription number 781058540]*
Publisher *Psychology Press*
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



The Quarterly Journal of Experimental Psychology Section A

Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713683590

Pupil Dilation as a Measure of Processing Load in Simultaneous Interpretation and Other Language Tasks

Jukka Hyönä [a]; Jorma Tommola [a]; Anna-Mari Alaja [a]
[a] University of Turku, Turku, Finland

Online Publication Date: 01 August 1995

To cite this Article Hyönä, Jukka, Tommola, Jorma and Alaja, Anna-Mari(1995)'Pupil Dilation as a Measure of Processing Load in Simultaneous Interpretation and Other Language Tasks',The Quarterly Journal of Experimental Psychology Section A,48:3,598 — 612

To link to this Article: DOI: 10.1080/14640749508401407

URL: http://dx.doi.org/10.1080/14640749508401407

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 1995, *48A* (3), 598–612

Pupil Dilation as a Measure of Processing Load in Simultaneous Interpretation and Other Language Tasks

Jukka Hyönä, Jorma Tommola, and Anna-Mari Alaja
University of Turku, Turku, Finland

The present study tested whether the pupillary response can be applied to study the variation in processing load during simultaneous interpretation. In Experiment 1, the global processing load in simultaneous interpretation as reflected in the average pupil size was compared to that in two other language tasks, listening to and repeating back an auditorily presented text. Experiment 1 showed clear differences between the experimental tasks. In Experiment 2, the task effect was replicated using single words as stimuli. Experiment 2 showed that momentary variations in processing load during a lexical translation task are reflected in pupil size. Words that were chosen to be more difficult to translate induced higher levels of pupil dilation than did easily translatable words. Moreover, repeating back words in a non-native language was accompanied by increased pupil dilations, in comparison to repetition in the subject's native language. In sum, the study lends good support to the use of the pupillary response as an indicator of processing load.

The study of cognitive demands during different language processing tasks has gained increased popularity as methods have been invented that yield an on-line record of cognitive processing. For example, reading research has gained new theoretical insights through the application of eye movement recordings (see Rayner & Pollatsek, 1989). The present study focuses on the pupillary response as an independent on-line measure of cognitive load, particularly during simultaneous interpretation and lexical translation. Simultaneous interpretation is known as a cognitively highly demanding task, which imposes heavy constraints on the interpreter's limited-capacity processing system.

The pupillary response has long been known to be associated with increased mental activity. Studies conducted during the last 30 years or so have provided evidence that human cognitive processes, such as problem solving or language comprehension, are accompanied with pupillary dilations (cf. reviews by Beatty, 1982; Janisse, 1977). After

---

Requests for reprints should be sent to Jukka Hyönä, Experimental Psychology Laboratory, Department of Psychology, University of Turku, Lemminkäisenk. 14-18 B, FIN-20520 Turku, Finland. Email: HYONA@sara.utu.fi.

This study was financially supported by the Turku University Foundation (Turun Yliopistosäätiö). We thank the two anonymous reviewers for their highly useful criticism. We are also grateful to Dick Olson for his comments on an earlier draft of this paper.

the seminal work by Kahneman and Hess (Hess & Polt, 1964; Kahneman & Beatty, 1966), pupillometry gained increased popularity in the 1970s in the study of momentary variations in the processing load during the performance of cognitive tasks, such as mental arithmetic (e.g. Ahern & Beatty, 1979), short-term memory tasks (e.g. Kahneman, Onuska, & Wolman, 1968), and language processing (e.g. Wright & Kahneman, 1971). Recently, the technique seems to have lost its appeal, as very few studies in which the pupillary response has been used as an indicator of processing load have been reported during the last decade. A literature search in PsychLiT yielded no more than five studies conducted during the past six years in which pupillary response was correlated with mental load (Backs & Walrath, 1992; Hoeks & Levelt, 1993; Just & Carpenter, 1993; Matthews, Middleton, Gilmartin, & Bullimore, 1991; Richer & Beatty, 1987). This is quite surprising, as the pioneering work seemed very promising. It is even more surprising considering the fact that recent advances in measurement technology have made the continuous recording of pupil diameter totally automated and clearly less cumbersome than when the pioneering work was conducted. Anecdotal evidence suggests that this decreased interest is due to researchers not being able to replicate some of the earlier findings. However, we could not find any reported replication failures in the literature. Whatever the reason for the decreased interest may be, we hope in the present study to bring back at least some of the enthusiasm that was once associated with the pupillometric method by showing its usefulness in the study of cognitive load during language processing.

The present study addresses the question whether the pupillary response is sensitive to variations in processing load during different language processing tasks, particularly during simultaneous interpretation and lexical translation. Simultaneous interpretation is a highly complex skill in which language perception, comprehension, translation, and production operations are carried out virtually in parallel and under severe time pressure. It is clear that such a complex activity is possible only by virtue of the independence and automaticity of many of the component processes. Yet the task is likely to create a heavy processing load. The sources and variations in mental load during interpretation present a question of theoretical and empirical interest.

In simultaneous interpretation, the interpreter is commonly found to lag well behind the speaker (lags of about 2–3 sec are typical during interpreting between two Indo-European languages—see Barik, 1973). Much longer lags are often necessary when the formal features of the input and output languages differ more extensively, as in the case of English and Finnish. This is because interpreting cannot be performed as simple lexical translation. The interpreter must frequently hold large segments of the source language utterance in abeyance, analysing the meaning, while simultaneously producing a different sentence in the target language. Consequently, the task heavily taxes the processing resources that the interpreter has at hand for information storage and computational operations. These mental processes are assumed to be performed in a limited-capacity working memory (e.g. Baddeley, 1986; Just & Carpenter, 1992). According to the working memory model proposed by Just and Carpenter (1992), there is a tradeoff between storage and computational demands, so that the more resources are allocated for one component, the less is left over for the other, provided that the processing needs exhaust the capacity resources. In a highly demanding task, such as simultaneous interpretation, such tradeoff seems quite probable.

Methodologically, the investigation of mental load during language processing requires a technique with which one can monitor the interpreter's performance on line. In other words, the technique should be sensitive to the process as it evolves in real time. A number of such on-line methods have been developed during recent years for the purposes of psycholinguistic research (cf. Carlson & Tanenhaus, 1989), but many of these involve subsidiary tasks that affect the actual language processing under study. The focus of the two experiments reported in this paper is a psychophysiological response—pupil dilation—which seems to offer some promise for on-line investigations of interpreting processes (see Tommola & Niemi, 1986). The measurement of task load through the pupillary response has the advantage that the researcher does not have to tamper with language processes themselves. So far it seems that no previous studies have applied the method to the study of cognitive processes in interpreting. Research on the validity of the technique as applied to interpreting situations thus appears well justified.

In the present study, Experiment 1 was conducted to see whether simultaneous interpretation would produce a greater increase generally in pupil size than would other, cognitively less demanding language processing tasks. In Experiment 1, the average pupil size during simultaneous interpretation of auditorily presented text passages was contrasted to listening to and shadowing (i.e. repeating back) a message. Experiment 2 employed single words instead of text passages, and the subjects listened to, repeated back, and orally translated the lexical stimuli. In addition to the comparison of pupil dilations across tasks, within-task variation in processing requirements was introduced in Experiment 2 by selecting two types of target words: words that were easily translatable and words that were more difficult to translate.

EXPERIMENT 1

Experiment 1 investigated the sensitivity of the pupillary response to reflect variations in processing load during language processing by comparing three different tasks that are self-evidently different in complexity. The tasks are (1) listening to a text passage, without any subsequent comprehension testing, (2) speech shadowing—that is, repeating back a message in the same language while listening to it, and (3) simultaneous interpretation into the subject's native language. The input language was English, and the output language in interpreting was Finnish. If pupillary dilation adequately reflects average mental load, the simple listening task should be associated with the lowest dilation levels, simultaneous interpretation should be associated with the highest dilation levels, and the shadowing task should fall between the two.

Method

Subjects. Nine subjects participated in Experiment 1. All were 3rd or 4th-year students at the University of Turku Department of Translation Studies, and all had Finnish as their native language. The subjects were highly fluent speakers of English, and all had received a minimum of one year's instruction in the techniques of simultaneous interpretation as one part of their study program.

Apparatus. The size of the pupil diameter was measured using an Applied Science Laboratories eye-tracker Model 1994. The computer connected to the eye-tracker was programmed to record the diameter of the subject's left pupil at a frequency of 50 Hz—that is, every 20 msec. The pupil data were temporally synchronized with the stimulus text by connecting the computer to the tape recorder that presented the source text through the subject's headphones and also preserved the subject's output in the interpreting and shadowing tasks.

Materials and Design. Three English-language texts (500–600 words) were employed as source texts. The materials had been written to simulate opening presentations at an international conference. Each task thus involved the comprehension of the non-dominant language (English). The shadowing task also involved speech production in English, whereas in simultaneous interpretation production was in Finnish. This change in the language of speech production works to increase the conservativeness of the design, as the supposedly more complex interpreting task involves production in the language the subjects had mastered best.

The texts were recorded by a native speaker of English. Three stimulus tapes, in which the order of presentation of the three passages was systematically varied, were constructed from the recordings. The nine subjects were assigned randomly to one of the three versions of the stimulus tape, three subjects per version. Task order was always the same: listening, followed by shadowing and interpreting. If there were a practice effect, the task performed last, simultaneous interpretation, would benefit most (i.e. the practice effect would decrease the pupil dilation especially for simultaneous interpretation), which would again increase the conservativeness of the design.

Thus, for each task, the design allowed an estimate of average processing load to be obtained on the basis of three texts, each of which was processed by three subjects.

Procedure. The subject was seated facing a computer screen located at a distance of approximately three metres. The room was illuminated with fluorescent lamps; no daylight was allowed to enter the room. To obtain a clear picture of the subject's pupil throughout the experiment and to prevent momentary variations in the amount of light reflected to the pupil, the subject was instructed to fixate his/her gaze on a graphic pattern shown on the screen during the performance of the tasks.

In the listening task, subjects were instructed to listen to each text passage with the intention of comprehending. However, no comprehension test was given afterwards. In shadowing, subjects were asked to repeat back the text in English at the same time as they were listening to it. In interpreting, they were instructed to translate the text simultaneously into Finnish. If they were not able to give an interpretation to some text segment quickly enough, they were instructed to leave that segment uninterpreted and to pick up the message again at the first possible point. This happened very rarely, however; the content of the text was not technical, and the interpreting skills of the trainees were already at a reasonably high level. The stimulus texts were presented to the subjects binaurally through headphones.

Three short warm-up texts preceded the actual experiment; during these, the subject practised each of the three tasks in the same order as that followed in the experiment proper. A 30-sec silence preceded and followed each practice text and experimental text. The total duration of subject performance (3 warm-up texts and 3 experimental texts) was approximately 19 minutes.

Results

Prior to analysis, all eye-blinks were removed from the data. To detect an eye-blink, an algorithm was devised that looked for abrupt drops in pupil diameter. An eye-blink was detected when the difference in pupil diameter between three consecutive samples exceeded a limit that was specified empirically. Secondly, the data were compressed into mean pupil size values corresponding to 3 sec of the performance. Such a mean value for a 3-sec window was calculated for every running second of the performance. The transformed data were fed into a spreadsheet program for the calculation of various statistics. The statistical analysis was performed using the *BMDP Statistical Software* (1990).

Table 1 gives the average pupil dilation level for the three tasks. It is apparent that the pupillometric estimate of mental load is consistent with the hypothesized difference in task difficulty. The analysis of variance computed from the data yielded a highly significant main effect of task, $F(2, 16) = 21.79$, $MS_e = 0.322$, $p < 0.001$. An analysis of contrasts showed that listening differed significantly from shadowing and shadowing from interpreting, both ps < 0.01. The contrasts computed by BMDP are similar to two-way t-tests.

To obtain an estimate of the task-internal variation in pupil size, each stimulus text was divided into three temporally equally long sections: beginning, middle, and end. The average pupil size for each section is given in Table 1 separately for each task. As may be seen in the table, the pupil dilation level was highest at the beginning of the task, after which it decreased noticeably. This effect was statistically significant, $F(2, 16) = 10.87$, $MS_e = 0.081$, $p = 0.001$. A post-hoc contrast yielded a reliable difference in pupil size between the beginning third and the middle third of the text passages, $F(1, 8) = 16.13$, $p < 0.01$. We suggest that this is a familiarity effect: the subject becomes acquainted with the conceptual content and other properties of the texts, as well as with the task itself. The means suggest that such a familiarity or practice effect may not be as apparent in interpreting as in the other two tasks. However, the Task × Time Sequence interaction failed to reach significance, $p = 0.12$.

According to an alternative interpretation, the decrease in pupil size during the task performance could be due to some sort of long-term autonomous drift in pupil size. To address the issue, we measured the pupil size for 30 sec immediately before and after each experimental task. For listening, the average pupil diameter was 4.74 mm before the experiment and 3.85 mm after it; for shadowing, the respective means were 4.55 mm and 4.47 mm, and for interpreting 4.92 mm and 4.88 mm. Thus, there does not seem

TABLE 1
Average Pupil Diameter in Experiment 1 as a Function of Task and Time Sequence

	Time Sequence			
Task	*Beginning*	*Middle*	*End*	*Mean*
listening	4.49	4.08	4.03	4.20
shadowing	4.98	4.59	4.58	4.72
interpreting	5.29	5.18	5.20	5.22
mean	4.92	4.62	4.60	4.71

Note: Pupil diameters are given in mm.

to be any noticeable drift in pupil size for shadowing or interpreting as the baseline measurements before and after the task yielded comparable pupil sizes. On the other hand, in listening the average pupil size was clearly larger before the task than after it. We believe, however, that this was not a drift effect, but arose from the degree of general alertness before and after the task performance. Subjects were probably more alert while waiting for the experiment to begin than after the easy listening task that did not require any overt response.

As is evident from Table 1, the difference in pupil diameter between the tasks remained throughout the task performance. Moreover, an analysis of the practice texts indicated that the task effect was not restricted only to the three experimental texts, but also emerged with the shorter texts the subjects processed prior to the experiment proper. Finally, there was also a carry-over effect: For average pupil size measured immediately after each task there was still a reliable difference between the tasks, $F(2, 16) = 18.41$, $MS_e = 0.134$, $p = 0.0001$ (see above for means). An analysis of contrasts revealed that listening produced a reliably smaller after-task pupil size than did shadowing, and shadowing produced a smaller pupil size than did interpreting, $p < 0.025$. An ANOVA on before-task pupil sizes yielded a main effect of task, $F(2, 16) = 6.01$, $MS_e = 0.780$, $p = 0.01$, which was due to shadowing having smaller average before-task pupil size than did interpreting, $p = 0.01$; other contrasts were non-significant.

Discussion

Experiment 1 indicated that the pupil record accurately reflects hypothesized differences in the average complexity of the three language processing tasks. Simultaneous interpretation clearly produced a higher degree of pupil dilation than did speech shadowing, and shadowing yielded a higher level than did listening. The technique seems therefore to hold some promise for studies that aim to describe how the interpreter utilizes his/her processing resources, and what factors affect the real-time performance.

The task differences we observed fulfil Kahneman's (1973) first criterion for a valid physiological indicator of processing load: Any indicator should reflect between-task differences elicited by qualitatively different cognitive operations (see also Beatty, 1982). However, there are two factors that may challenge the validity of Experiment 1: (1) On the basis of the global task effect observed in Experiment 1, one cannot draw any conclusions as to whether the pupillary response reflects momentary fluctuations in processing load. The task effect may merely reflect a long-term increase in the level of general arousal. The carry-over effect of task difficulty after the completion of the task proper supports such a conclusion. (2) The fact that listening induced the lowest levels of pupil dilation may also be attributed to the lack of any output requirements, which were an integral part of the other language tasks. These two concerns were addressed in Experiment 2.

EXPERIMENT 2

Experiment 2 was conducted to tackle the two confounding factors in Experiment 1. It was performed using single words as stimuli instead of whole text passages. This allowed us to register the changes in the pupil size in a more moment-to-moment fashion than in

Experiment 1, which was based on global analyses only. Thus, the aim was to provide evidence for the second criterion proposed by Kahneman (1973) for a physiological indicator of processing load—namely, that it should also be sensitive to within-task variations in task demands, in addition to reflecting qualitative differences between tasks.

The language processing tasks in Experiment 2 involved the following modifications. An output requirement was also included in the listening task. In listening, subjects were required to say "yes" aloud after they had recognized the stimulus word. In shadowing they were to repeat the word aloud. The simultaneous interpretation used in Experiment 1 was replaced with a lexical translation task in which subjects were to give a meaning equivalent for the word in the output language. Lexical translation is a component process of simultaneous interpretation; however, it should be kept in mind, as noted earlier, that simultaneous interpretation also involves other cognitive operations.

In Experiment 2 we also varied the input language by having both English and Finnish as the source language. The idea was to see whether a non-native language as the source language would lead to more effortful processing and higher levels of pupil dilation.

To vary the task demands within the same task, we introduced two types of input words: those for which it was either relatively easy to find a one-to-one equivalent in the output language and those for which this was more difficult. If the relative pupil size reflected variations in within-task demands, this effect should be observable in the lexical translation task.

Method

Subjects. Eighteen new subjects meeting the same criteria as in Experiment 1 participated in Experiment 2. One subject had to be dropped from subsequent analyses because of equipment failure.

Materials. To create two sets of words that would differ in their relative translatability, we chose 125 English and 145 Finnish words using the following criteria: For more difficult items, words were selected that did not have an obvious one-word equivalent in the other language but had to be translated by a phrase (e.g. "undertake" = "*ottaa tehtäväkseen*"; "*harrastaa*" = "have something as a hobby"). For easy lexical items, on the other hand, words were selected that could be translated into the other language using a single word (e.g. "journal" = "*aikakausilehti*"; "*korostaa*" = "emphasize"). The English words represented the absolute frequency ranks of 13–21 in the LOB Corpus rank list (Hofland & Johansson, 1982). The Finnish words were taken from the frequency range of 20–49 in the Saukkonen, Haipus, Niemikorpi, and Sulkala (1979) corpus. The words were from different grammatical categories.

The final selection of the target words was made on the basis of a pretest. Fourteen first- and second-year students from the School of Translation Studies participated in the pretest; all were native speakers of Finnish but were fluent in English. The words were presented to the subjects auditorily. English and Finnish words were presented in separate blocks, and the order of language was counterbalanced across subjects. Words were presented at 5-sec intervals, with items that had originally been judged as "easy" alternating with those judged "difficult" in blocks of ten words. The subjects were to translate each word as soon and as accurately as possible. All responses were registered on tape for further analysis. Subjects were not told of the presumed differences in the translatability of the words beforehand, nor were there pauses between the blocks indicating block change. The two language blocks were separated by a pause of 30 sec. Prior to both language blocks, 5 practice words were presented. The responses were analyzed for correctness of translation. On the basis of the analysis, a translatability rank list was created for words in both languages, where words were listed according to the mean probability of correct translation. On the basis of the rank list, 90 words per language were chosen for the actual experiment. The set of easy items included the first 45 words from the top of the rank list; difficult items consisted of another 45 words from the bottom of the list. Of the 45 words in each set, one third (i.e. 15 words) were presented in each of the three tasks of Experiment 2, so that, across subjects, each word appeared equally often in each task.

Procedure. The same apparatus was used as in Experiment 1. In each task, the stimulus words were presented binaurally one word at a time through headphones. The Finnish and English words were presented in separate blocks, and the order of language was counterbalanced across subjects. As in Experiment 1, the three experimental tasks were blocked, and their order was counterbalanced across subjects. The words were presented at 10-sec intervals. One second prior to each word a buzzer signal was given to alert the subject that a word would be presented. Subjects were instructed to react to the word as soon as possible, according to the task requirements: when listening, they were to say "*on*" ["yes"], when shadowing they were to repeat the word, and for the lexical translation task they were to translate the word into the other language. Subjects' responses were taped on audio cassettes. Within a block, difficult and easy words alternated so that every second word was difficult.[1] Prior to each task block, 5 practice words were presented. After each block, subjects were allowed to rest for a couple of minutes.

Results

Pupil Data. The analysis of the raw pupil data was carried out in the following way. As individual tapes were used for stimulus presentation instead of one master tape, possible differences in exact presentation times were checked. This was done by recording the stimulus material and the subject's output (recorded on a separate track of the stimulus presentation tape) via MacRecorder to the Macintosh SoundEdit application, which produced a graphic presentation of the sound wave together with a time scale. This procedure was used to determine the exact time of the onset of the buzzer signal, the stimulus word, and the subject's response for each word. The pupil data were analysed using a program that drew a graphic presentation of all the data points (i.e. 50 per sec) together with a time scale. For each stimulus word, mean and maximum pupillary values were measured during two time periods: the baseline period from 4 to 3 sec before the buzzer signal, and the task performance period from stimulus onset to the beginning of the output. The peak amplitude of pupil dilation typically occurs approximately 900–1200 msec after the beginning of an experimental trial, and dilation begins with a delay of 300–500 msec (see Beatty, 1982; Hoeks & Levelt, 1993). Our task performance period ranged from 1100 msec to 2500 msec. Thus in the majority of cases the peak amplitude occurred within our recording interval. Recording was terminated at the onset of the

---

[1] It may be argued that the alternation between easy and difficult words would make subjects aware of the fact that every second word required a phrase as a response. This is not very likely, however, because quite a few of the easy words translated into compounds that could be construed as two separate words.

response, because the pupil starts constricting in the verbal output phase (see Beatty, 1982).

Below we report ANOVAs for the average increase in pupil size during the task performance in comparison to the average baseline level. The results were very similar when the increase in the maximum pupil size was used as the dependent variable. When computing the average pupil size, all data associated with eye blinks were excluded using the same procedure as in Experiment 1.

The experiment had a 2 × 3 × 2 within-subject design. The factors were input language (English, Finnish), task (listen, shadow, interpret), and word translatability (easy, difficult). ANOVAs were conducted using both subjects (F_1) and stimulus items (F_2) as the random factor. In Table 2, the average increase in pupil size is presented for each condition, together with the mean pupil size during the baseline and performance phase.

The main effect of task was highly significant, $F_1(2, 32) = 20.38$, $MS_e = 0.024$, $p < 0.0001$, $F_2(2, 352) = 60.75$, $MS_e = 0.025$, $p < 0.0001$. The overall average increase in pupil size was largest for lexical translation (0.33 mm), followed by shadowing (0.24 mm), and by listening (0.16 mm). All the pairwise contrasts between tasks were significant, all ps < 0.01. The values for maximum pupil dilation showed an analogous pattern: The maximum dilations for lexical translation, shadowing, and listening were 0.47 mm, 0.33 mm, and 0.20 mm, respectively.

The main effect of language proved significant, $F_1(1, 16) = 4.89$, $MS_e = 0.034$, $p < 0.05$, $F_2(1, 176) = 12.84$, $MS_e = 0.024$, $p < 0.001$. The pupil dilated more for the English words (0.27 mm) than for the Finnish words (0.22 mm). The main effect of word translatability was also highly significant, $F_1(1, 16) = 32.89$, $MS_e = 0.012$, $p < 0.0001$, $F_2(1, 176) = 48.87$, $MS_e = 0.024$, $p < 0.0001$. More difficult words produced a larger increase in pupil size (0.29 mm) than did the easily translatable items (0.20 mm).

TABLE 2
Average Pupil Diameter in Experiment 2 during Baseline and Performance, as a Function of Language, Task, and Word Translatability

	Word Translatability					
	Easy			*Difficult*		
Task	*Baseline*	*Performance*	*Diff.*	*Baseline*	*Performance*	*Diff.*
Finnish						
listening	5.39	5.51	0.12	5.38	5.57	0.19
shadowing	5.40	5.56	0.16	5.41	5.62	0.21
translating	5.76	5.98	0.22	5.64	6.04	0.40
English						
listening	5.29	5.44	0.15	5.32	5.50	0.18
shadowing	5.48	5.75	0.27	5.47	5.81	0.34
translating	5.81	6.10	0.29	5.71	6.12	0.41

Note: Pupil diameters are given in mm.

The main effects were qualified by two interactions. The Task × Translatability interaction, $F_1(2, 32) = 5.51$, $MS_e = 0.008$, $p < 0.01$, $F_2(2, 352) = 8.14$, $MS_e = 0.025$, $p < 0.001$, suggests that in lexical translation the pupil clearly dilated more for difficult words than for easy words, with average dilations of 0.40 mm (difficult words) and 0.25 mm (easy words). In listening and shadowing the same trend was observable, but to a much lesser degree. In listening, the average pupil dilation was 0.19 mm for the difficult words and 0.13 mm for the easy words; in shadowing the respective means were 0.27 mm and 0.21 mm. A similar pattern was observed for maximum pupil dilation: In translation the pupil dilation was clearly greater for difficult words (0.57 mm) than for easy words (0.37 mm); in shadowing the same pattern was visible, although the difference between difficult and easy words was not as pronounced (0.37 mm vs. 0.28 mm), the difference being even smaller in listening (0.23 mm vs. 0.19 mm).

The Task × Language interaction proved significant by items, $F_2(2, 352) = 6.65$, $MS_e = 0.025$, $p < 0.01$, but was statistically marginal by subjects, $F_1(2, 32) = 2.60$, $MS_e = 0.022$, $p < 0.1$. An analysis of simple effects showed that in shadowing the relative difference in pupil size between the baseline and the task performance condition was larger in English than in Finnish, $F_1(1, 16) = 6.23$, $p < 0.05$, $F_2(1, 176) = 6.12$, $p = 0.01$, but this was not the case in listening or in lexical translation.

We also computed the probability of making an eye-blink during the baseline and task performance phase. Subjects tended to blink more in the lexical translation task (the mean was 0.64) than during the other tasks (the means were 0.39 and 0.42, for listening and shadowing, respectively), $F(2, 32) = 11.49$, $MS_e = 0.224$, $p < 0.001$. The difference was particularly salient during task performance, where the mean number of eye-blinks was 0.45 for listening, 0.48 for shadowing, and 0.86 for lexical translation. During the baseline period the difference was not as evident (the respective means were 0.33, 0.36, and 0.43). The reliability of this pattern was confirmed by a significant Task × Time Sequence (baseline vs. performance) interaction, $F(2, 32) = 10.95$, $MS_e = 0.105$, $p < 0.001$. This result suggests that even the blinking rate may be sensitive to cognitive load, so that heavier load is associated with a higher blinking rate. This runs counter to prior research that has shown an inverse relationship between blinking rate and task difficulty (Bagley & Manelis, 1979; Holland & Tarlow, 1972). The present data are not totally conclusive, however, as the difference in blinking rate is confounded with somewhat longer recording times for lexical translation due to longer response latencies (see below).

Response Latencies. The response latency was the time that elapsed from stimulus offset to the onset of the required verbal response. The response latencies were analysed using three within-subject factors: language, task, and word translatability. The average response latencies for each experimental condition are given in Table 3.

All three main effects proved significant. The main effect of language, $F_1(1, 16) = 10.33$, $MS_e = 45{,}941$, $p < 0.01$, $F_2(1, 176) = 9.13$, $MS_e = 140{,}477$, $p < 0.01$, suggests longer overall response latencies for Finnish words (1530 msec) than for English words (1434 msec). The main effect of task, $F_1(2, 32) = 126.79$, $MS_e = 110{,}731$, $p < 0.0001$, $F_2(1, 176) = 342.43$, $MS_e = 116{,}559$, $p < 0.0001$, indicates that response latencies in lexical translation were remarkably longer (2002 msec) than in the listening (1279 msec) or in the shadowing (1164 msec) task. Response latencies in the three language tasks also

TABLE 3
Average Response Latencies for Target Words, as a Function of Language, Task, and Word Translatability

	Finnish		*English*	
Task	*Easy*	*Difficult*	*Easy*	*Difficult*
listening	1296	1300	1212	1308
shadowing	1185	1216	1090	1167
translating	1683	2502	1659	2166
mean	1388	1673	1320	1547

Note: Response latencies are given in msec.

differed from each other separately, all ps < 0.0001. The main effect of word translatability, $F_1(1, 16) = 59.13, MS_e = 56{,}509, p < 0.0001, F(1, 176) = 78.99, MS_e = 140{,}477, p < 0.0001$, suggests longer response latencies for difficult words (1610 msec) than for easy words (1354 msec).

The main effects were qualified by two interactions, Task × Word Translatability $F_1(2, 32) = 37.88$, $MS_e = 55{,}794$, $p < 0.0001$, $F_2(1, 176) = 61.68$, $MS_e = 166{,}559$, $p < 0.0001$, and Language × Task × Word Translatability $F_1(2, 32) = 6.89$, $MS_e = 30{,}050$, $p < 0.01$, $F_2(1, 176) = 8.11$, $MS_e = 116{,}559$, $p < 0.001$. The former interaction indicates that it was particularly in lexical translation that the response latencies for difficult words were markedly longer than for easy words. The three-way interaction was partialled out by analyses of simple effects. We consider first the results for Finnish stimulus words. It was found that in listening the difficult words did not produce longer response latencies than did easy words, Fs < 1. In shadowing, the trend seemed reliable by subjects, $F_1(1, 16) = 4.97, p < 0.05$, but not by items, $F_2(1, 176) = 1.15, p > 0.1$. In lexical translation, difficult words produced markedly longer response times than did easy words, $F_1(1, 16) = 28.26$, $p = 0.0001$, $F_2(1, 176) = 63.90$, $p < 0.0001$. On the other hand, difficult English words produced longer latencies in all three tasks: in listening, $F_1(1, 16) = 60.39$, $p < 0.0001$, $F_2(1, 176) = 7.71$, $p < 0.01$; in shadowing, $F_1(1, 16) = 48.78$, $p < 0.0001$, $F_2(1, 176) = 8.05$, $p < 0.01$; as well as in lexical translation, $F_1(1, 16) = 40.28$, $p < 0.0001$, $F_2(1, 176) = 19.02$, $p < 0.0001$.

Discussion

The results of Experiment 2 proved very encouraging with respect to the use of pupil dilation as an indicator of moment-to-moment variation in cognitive load during language processing. Pupil diameter reflected variation in processing requirements within task as well as between tasks. It is noteworthy that most of the effects were also very robust in the sense of statistical significance.

The results can be summarized in three main points.

(1) The pupillary response was shown to vary as a function of task difficulty. The increase was largest when the task required a translation of a word from one language to

another, it was smaller when the task was to repeat a word aloud, and it was smallest when subjects simply had to acknowledge that a word had been presented auditorily. It should be noted that the increase in pupil size from the baseline level was greatest in lexical translation, despite the fact that the pre-trial baseline level was considerably higher for lexical translation (5.73 mm) than for shadowing (5.44 mm) or for listening (5.35 mm). This compares favourably with earlier studies that showed that the magnitude of the pupillary response is independent of the baseline level over a range of values that are not extreme (see Beatty, 1982). Our three language tasks are an addition to the summary provided by Beatty (1982) of the peak amplitudes of pupillary response in different cognitive tasks (see his Figure 8). Beatty's summary indicated systematic variation in the peak amplitudes due to task difficulty; for instance, a perception task induces lower levels of pupil dilation than, say, a reasoning task. The replication of the task effect in Experiment 2 rules out the possibility that the difference observed in Experiment 1 between listening and the other two tasks was primarily due to differences in the immediate output requirements.

(2) In lexical translation, words that were difficult to translate into the output language brought about a larger increase in pupil size from the baseline level than did easily translatable words. This is the most direct piece of evidence in support of the view that momentary increase in processing load is reflected in higher levels of pupil dilation. Interestingly, the translatability effect also seemed to be observable in shadowing and listening, although to a lesser extent. This suggests that the lexical item in the other language seems to become activated even though the task does not require an overt translation. This can be taken as evidence for a theory that among bilinguals the native language and the second language are in close contact with each other. This notion was further supported by the latency data, which demonstrated longer latencies for difficult English words even in listening and shadowing.

The average increase in pupil dilation during the translation of difficult words was 0.40 mm, and the maximum dilation was 0.57 mm. In Beatty's review, the highest peak amplitudes of pupillary response were observed in mental arithmetic—that is, about 0.5 mm for the most difficult multiplication problems. Thus, our peak amplitude appears to be the highest reported in the literature.

(3) The pupil dilated more among the bilingual Finnish subjects when they repeated English words than when they repeated Finnish words. On the assumption that it is more cognitively demanding to repeat words in a non-native language, this is further evidence for the notion that the pupillary response is sensitive to cognitive load.

The results obtained using the pupillary response as an indicator of the difficulty of processing were generally replicated by the latency data of the verbal responses. (1) Lexical translation clearly produced longer response latencies than did the other tasks. In contrast to the pupil data, however, shadowing produced somewhat shorter latencies than did listening, probably because the listening task did not encourage a speeded response to the same extent as did shadowing. (2) For lexical translation, response latencies appeared to be longer for difficult words than for easy words. This was true for both input languages. Moreover, when English was used as the source language, response latencies seemed to be longer for difficult words not only in translation but also in the other two tasks (a plausible interpretation of the effect was given earlier).

Although the differences found between the experimental tasks in Experiment 1 could perhaps be interpreted as a long-term emotional arousal due to stress, it is very unlikely that the same argument could apply to the results of Experiment 2. Experiment 2 showed that the pupil dilated during task performance in comparison to the baseline level, and it dilated more with more difficult tasks. After a response had been given to a target word, the pupil constricted back to the baseline level. In other words, there was considerable moment-to-moment variation in pupil size within each task. It is not likely that this kind of fast phasic fluctuation in the pupil size could be attributed to a general emotional arousal. As Beatty (1982) has pointed out, "the effects of emotional arousal are generally longer lasting than the brief phasic responses evoked by cognitive activity" (p. 288). Thus, relative changes in pupil size due to variation in task demands within a same task are difficult to interpret as a long-term emotional effect. It should be noted, however, that in addition to the fast phasic fluctuation in Experiment 2, and to some extent in Experiment 1 also, an increase in the level of dilation was observed as a carry-over effect: The baseline level of the pupil, as measured between stimuli, was, on the average, larger in translation (5.73 mm) than in shadowing (5.44 mm) or in listening (5.35 mm). Thus, superimposed on the task-induced moment-to-moment variation in pupil size, there also seems to be more durable effect, possibly caused by variation in the general arousal level.

The results from lexical translation cannot be directly applied to simultaneous interpretation. As was noted earlier, simultaneous interpretation cannot be construed as a process of simple lexical look-up, as the interpreter has to carry out other cognitive operations that go well beyond lexical translation. Thus, the interpreter cannot translate the input language one word a time (for reasons discussed above) but has to take in and hold larger chunks of information before a translation can be carried out. This means, among other things, that simultaneous interpretation involves temporary storage of information, which is relatively absent in lexical translation. Nevertheless, Experiment 2 demonstrated that lexical translation was coupled with increased cognitive load. Thus, it can be concluded that the computational requirements themselves, apart from temporary storage of information, increase the processing load and tax the limited processing resources the interpreter has available. It is left for future research to evaluate the independent contributions of storage and computational requirements in creating the total processing load associated with simultaneous interpretation.

GENERAL DISCUSSION

The present study was designed to test whether the technique of pupillometry could be applied successfully to study variations in processing load during language processing, particularly during simultaneous interpretation and lexical translation. The two experiments provided consistent evidence for the view that the pupillary response reflects qualitative differences between tasks, and Experiment 2 found good support for the view that moment-to-moment variation in processing load during lexical translation can also be observed in the extent of pupil dilation.

Kahneman (1973) proposed three criteria for any physiological indicator of processing load: It should capture (1) differences in processing load between different cognitive tasks, (2) variations in processing load within a same task, as well as (3) individual differences in the ability to carry out a cognitive task. The present study provided evidence for the first two criteria. We are currently planning a study that would bear on Kahneman's third criterion by comparing the extent of pupil dilation of practised interpreters with those of novice interpreters. To date, there is some evidence that individual differences may be reflected in the pupillary response. Ahern and Beatty (1979) reported higher levels of pupil dilation for individuals with low IQs than for those with high IQs while doing a mental arithmetic task. Antikainen and Niemi (1983) showed that neurotic subjects produce a stronger pupillary response to noise than do less neurotic subjects. On the other hand, Just and Carpenter (1993) were not able to show differences in pupil size during sentence processing among subjects with a low versus a high working memory span.

The study of cognitive load during language processing bears both theoretical and practical significance. Therefore, it is important to have independent and reliable measures of processing load. The present study demonstrates that the pupillary response can be successfully employed as an on-line measure to reflect moment-to-moment changes in cognitive load during language processing. Consequently, the present study gives researchers licence to employ pupillometry in the study of more theoretically motivated questions.

REFERENCES

Ahern, S.K., & Beatty, J. [1979). Physiological signs of information processing vary with intelligence. *Science, 205*, 1289–1292.

Antikainen, J., & Niemi, P. [1983). Neuroticism and the pupillary response to a brief exposure to noise. *Biological Psychology, 17*, 131–135.

Backs, R.W., & Walrath, L.C. (1992). Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. *Applied Ergonomics, 23*, 243–254.

Baddeley, A.D. (1986). *Working memory.* Oxford: Oxford University Press.

Bagley, J., & Manelis, L. (1979). Effect of awareness on an indicator of cognitive load. *Perceptual and Motor Skills, 49*, 591–594.

Barik, H. (1973). Simultaneous interpretation: Temporal and quantitative data. *Language and Speech, 16*, 237–270.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*, 276–292.

BMDP Statistical Software (1990). Berkeley, CA: University of California Press.

Carlson, G.N., & Tanenhaus, M.K. (Eds.) (1989). *Linguistic structure in language processing.* Dordrecht: Kluwer.

Hess, E.H., & Polt, J.H. (1964). Pupil size in relation to mental activity during simple problem solving. *Science, 143*, 1190–1192.

Hoeks, B., & Levelt, W.J.M. (1993). Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research Methods, Instruments, & Computers, 25*, 16–26.

Hofland, K., & Johansson, S. (1982). *Word frequencies in British and American English.* Bergen: NAVF.

Holland, M.K., & Tarlow, G. (1972). Blinking and mental load. *Psychological Reports, 31*, 119–127.

Janisse, M.P. (1977). *Pupillometry: The psychology of the pupillary response.* Washington: Hemisphere Publishing.

Just, M.A., & Carpenter, P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122–149.

Just, M.A., & Carpenter, P.A. (1993). The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology, 47*, 310–339.

Kahneman, D. (1973). *Attention and effort.* Englewood Cliffs, NJ: Prentice-Hall.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science, 154*, 1583–1585.

Kahneman, D., Onuska, L., & Wolman, R.E. (1968). Effects of grouping of the pupillary response in a short-term memory task. *Quarterly Journal of Experimental Psychology, 20*, 309–311.

Matthews, G., Middleton, W., Gilmartin, B., & Bullimore, M.A. (1991). Pupillary diameter and cognitive load. *Journal of Psychophysiology, 5*, 265–271.

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading.* Englewood Cliffs, NJ: Prentice-Hall.

Richer, F., & Beatty, J. (1987). Contrasting effects of response uncertainty on the task-evoked pupillary response and reaction time. *Psychophysiology, 24*, 258–262.

Saukkonen, P., Haipus, M., Niemikorpi, A., & Sulkala, H. (1979). *Suomen kielen taajuussanasto* [Frequency Dictionary of Finnish]. Porvoo: WSOY.

Tommola, J., & Niemi, P. (1986). Mental load in simultaneous interpreting: An on-line pilot study. In L. Evensen (Ed.), *Nordic research in text linguistics and discourse analysis* (pp. 171–184). Trondheim: Tapir.

Wright, P., & Kahneman, D. (1971). Evidence for alternative strategies of sentence retention. *Quarterly Journal of Experimental Psychology, 23*, 197–213.

Revised manuscript received 30 August 1994