Behavior Research Methods
2008, 40 (4), 988-1000
doi: 10.3758/BRM.40.4.988

The use of Tholos software for combining measures of mental workload: Toward theoretical and methodological improvements

Julien Cegarra
Université de Toulouse, Toulouse, France

and

Aline Chevalier
Université Paris Ouest-Nanterre-La Défense, Paris, France

Mental workload is a concept central to a range of disciplines (including cognitive psychology and ergonomics) that has given rise to various theoretical and methodological debates. As a result, researchers have used a number of techniques for measuring mental workload. Traditionally, three categories of measurement technique have been recognized: performance measures (e.g., the dual-task paradigm), subjective measures (e.g., rating scales), and physiological measures (e.g., heart, respiration, and blink rates). Each technique has advantages and limitations; however, some limitations may prevent an accurate evaluation of the mental workload. In this article, we focus on the benefits of combining multiple measures of mental workload. However, because combining several techniques is a very complex process, we have developed the Tholos software in an attempt to reduce this complexity. This software package uses measures from each of the three categories: a dual-task paradigm with auditory signals; the National Aeronautics and Space Administration task load index (NASA-TLX) scale and its simplified version, the "raw" task load index (RTLX); and physiological (such as pupil-dilation) measurements with which our software can merge results from the dual-task paradigm. To illustrate the benefits of using Tholos, we describe a puzzle-solving experiment in which we combined multiple measures of mental workload. The results indicate the importance of combining multiple measures to build upon the theoretical and methodological foundations of mental workload.

The concept of *mental workload* (or *cognitive load*) is central to many disciplines, including cognitive psychology, educational psychology, and cognitive ergonomics. For example, in cognitive ergonomics, there is a need to regulate the human operator's mental workload in order to prevent both underload and overload, while maintaining an adequate performance level. At the same time, the diverse theoretical foundations of mental workload are still under debate, and there is no consensus about how the concept of mental workload should be defined. Basically, however, the notion presupposes that cognitive processes have a cost that draws from a finite pool of cognitive resources. Mental workload is therefore often defined as the ratio between the demands of the task and the human resources available.

We briefly describe the three main categories of mental workload measurements—performance-related, subjective, and physiological—after which, we survey the benefits and necessity of combining multiple methods to assess mental workload, first in terms of their requirements and then of the quantity and quality of the data obtained. We developed the Tholos[1] software to make it easier to combine multiple measures of mental workload. In an overview of the Tholos design goals and the measures implemented, we provide a case study involving puzzle solving to illustrate the benefits of using Tholos to combine measures. We conclude with a discussion of the theoretical and methodological gains achieved through combining multiple measures.

Three Categories of Mental Workload Measurement

Methods for measuring mental workload are usually classified using antitheses, such as *direct* versus *indirect*, *objective* versus *subjective*, or *analytical* versus *empirical*. Brünken, Plass, and Leutner (2003) suggested that methods can be classified along two of these antitheses: direct or indirect (*causal relation*) and objective or subjective (*objectivity*). One of the most intuitive classifications focuses on the means used to assess mental workload, taking into consideration *performance measures* (in the primary or additional task), *subjective measures* (e.g., rating scales), and *physiological measures* (e.g., heart rate, respiration rate, and blink rate). Many studies have fo-

J. Cegarra, julien.cegarra@univ-jfc.fr

cused on comparing methods belonging to one of these categories, especially subjective reports (e.g., Hill et al., 1992; Rubio, Díaz, Martín, & Puente, 2004) or physiological measures (e.g., Veltman & Gaillard, 1998). Miyake (2001) suggested using a "multivariate workload evaluation index" that combines physiological measures and subjective reports by means of principal-components analysis. Ryu and Myung (2005) studied the relationship between subjective measures (i.e., NASA-TLX, the National Aeronautics and Space Administration task load index) and various physiological measures. They noted that a regression equation including three physiological measures (heart rate variability, blink intervals, and alpha rhythm) accounted for 51% of the subjective score. Following this fruitful approach, as well as the literature on mental workload, we have selected a good example from each category and discuss how to combine methods from these categories.

Performance measures. Performance measures can be divided into measurements of the mental workload directly from the main task or from an additional task that is being performed concurrently.

Measurement of performance in the main task. Many studies indicate a strong link between poor performance in the main task and the mental workload experienced by the participant (e.g., Bi & Salvendy, 1994). The mental workload involved in a given task can be evaluated using a variety of performance metrics, such as time taken, speed, or number of errors. However, such an approach is not always reliable, and care is required when interpreting such measures of mental workload. For example, it is necessary to distinguish between low performance that is related to excessive resource requirements and low performance that is related to the quality of the data currently available (e.g., Norman & Bobrow, 1975). Furthermore, depending on several factors, two contrasting levels of performance (low and high) may result from either a low or a high mental workload. More precisely, the relationship between mental workload and performance depends on task demands, human strategies, motivation, and individual differences (e.g., Bainbridge, 1974; Cegarra & Hoc, 2006; Paas, Tuovinen, van Merriënboer, & Darabi, 2005).

Measurement of an additional task. The idea underlying measurement of an additional task is that the capacity that is not being used to perform the primary task can be used to perform another task. A common example of a combination of principal and additional tasks is driving while either holding a conversation or receiving calls on a mobile phone. The performance of the additional task can be used as an index of the demands made by the primary task (e.g., there may be gaps in the telephone conversation). The dual-task paradigm is the approach most frequently used for additional task measurement: The participants are asked to complete a task (considered to be the primary or main task), and at various times they are interrupted by a signal (e.g., a tone or a brief flash of light). They have to respond to this signal as quickly as possible—for example, by clicking with a mouse (this constitutes the additional or secondary task).[2] It is assumed that the time taken to respond to the signal (the additional task) reflects the amount of cognitive resources allocated to the primary task. Piolat, Olive, Roussey, Thunin, and Ziegler (1999) suggested selecting a probe from a different pool of resources in order to avoid interfering with the participant's main task. Although some researchers (e.g., Fisk, Derrick, & Schneider, 1987) have challenged the assumptions made by this method, many others (e.g., Chevalier & Kicka, 2006; Kellogg, 1990; Levy & Ransdell, 1995; Piolat et al., 1999) have used it successfully in a range of tasks, including design, text writing, and information searching.

The limitations of primary task measurement have been noted. The dual-task paradigm appears, therefore, to be the most relevant performance measure of mental workload in this category. Although using the dual-task paradigm is more artificial than using a familiar additional task (such as holding a conversation while driving), this paradigm makes it possible to compare the results from one study with those from another by using a common index. Basing our research on a validated dual-task paradigm may facilitate discussion about any concerns regarding this performance measure.

Subjective measures. Subjective measures consist of asking participants to provide judgments of the cognitive effort required to complete a task after it has been completed. Subjective reports may appear questionable because they are based on the assumption that participants are able to report accurately on their mental workload. However, they have several advantages: They are relatively nonintrusive with regard to the task carried out by the participant, they are easy to implement, and they have significant theoretical support (e.g., Hart & Staveland, 1988; James, Elderfield, Palmer, & Connelly, 1995). Moreover, participants are usually very consistent in their self-assessment of multiple ratings (Schvaneveldt, Reid, Gomez, & Rice, 1998). Several scales have been developed. Some of those used most often are SWAT (the subjective workload assessment technique), NASA-TLX, and MCH (the modified Cooper–Harper scale; Cooper & Harper, 1969), all of which are outlined below.

SWAT (Reid, Potter, & Bressler, 1987) uses three scales: *time load*, *mental effort load*, and *psychological stress load*. Each scale has three discrete levels: low, medium, and high. In a pretask procedure, the participant is given a set of 27 cards, each containing a pairwise comparison of the three scales at the three levels. The participant grades the mental workload, from the lowest to the highest. The order obtained reflects the participant's assessment of her/his own mental workload. After dealing with this pretask procedure, the participant then rates the three scales after each trial in the experiment. Several authors have pointed out that this pretask procedure is very time-consuming and have suggested that it could be replaced with a simpler comparison of pairs of scales, or even could be eliminated from the pretask procedure (Luximon & Goonetilleke, 2001; Moroney, Biers, & Eggemeier, 1995). Although SWAT has been used extensively, Hart and Staveland (1988) have found that it cannot discriminate between subtle differences at the lowest levels of task demand.

NASA-TLX (Hart & Staveland, 1988) is also a multidimensional rating scale procedure. It considers six dimensions of load assessment: *cognitive demand*, *physical demand*, *temporal demand*, *effort*, *performance*, and *frustration*. At the end of the main task, the participant scores each dimension from 0 to 100. These six dimensions are then displayed in pairs, and the participant selects the dimension that contributed most to her/his load. This posttask procedure makes it possible to assess the importance of these dimensions in the final calculation of the load. Because this second step is time-consuming, Byers, Bittner, and Hill (1989) suggested calculating an overall load RTLX; this is an average of the scores for the six dimensions, and so does not require the pairwise comparison of dimensions. They also demonstrated that the value obtained is closely correlated with the traditional calculation. This RTLX may therefore be useful, particularly in reducing the time taken to collect data, which is especially important in many field studies.

MCH (Wierwille & Casali, 1983) is a 10-point, one-dimensional rating scale based on a decision tree. It was developed for psychomotor tasks in order to rate aircraft handling and control (Cooper & Harper, 1969; see also Lysaght et al., 1989). Wierwille and Casali extended this scale to enable it to take a larger set of tasks into account. Because this is a one-dimensional scale, it has the advantage of permitting fewer interpretations of the ratings for each participant. Moreover, Hendy, Hamilton, and Landry (1993) noted that, generally, a one-dimensional scale can predict an overall load as well as a multidimensional scale can. In contrast to the SWAT and NASA-TLX scales, however, the valuable diagnostic information provided by a multidimensional rating is no longer available for analysis. Another shortcoming of the MCH scale is its lack of sensitivity in comparison with other scales, such as SWAT (e.g., Kilmer et al., 1988).

From these three scales, we can note that the NASA-TLX is often considered to be not only the most sensitive, but also the most reliable subjective measure (Hill et al., 1992). Furthermore, because NASA-TLX is a multidimensional rating, it also provides valuable diagnostic information about the source of the load (Hart & Staveland, 1988).

Physiological measures. Physiological measures—for example, of variations in heartbeat, pupil dilation, blink rate, or respiration rate—are based on the assertion that physiological variables reflect changes in cognitive functioning. They particularly require participants to report recent use of alcohol or caffeinated drinks, cigarettes, or drugs, whose use may alter physiological responses (e.g., Granholm, Asarnow, Sarkin, & Dykes, 1996). After such factors are accounted for, these measures are considered precise enough to assess subtle variations in mental workload over a continuous time frame.

However, the suitability of some physiological techniques for assessing mental workload has been challenged. For example, Lee and Park (1990) noted that changes in mental workload do not affect the heart rate, but, rather, the heart-rate variability (*sinus arrhythmia*). However, they also noted that an increase in the physical load modified both heart rate and heart-rate variability. This highlights the fact that, even after identifying structures relevant for measuring mental workload, one has to consider the ratio between the "signal" (produced by the relevant structure) and "noise" (due to unwanted structures). This signal-to-noise ratio has been the topic of many studies of physiological measures of mental workload, especially electroencephalography (Wilson & O'Donnell, 1988). The effect of noise indicates potential limitations of physiological measures, which include changes not related solely to mental workload (i.e., a lack of selectivity).

Wickens and Hollands (2000) stated that physiological measures have to be considered within the context of separate pools of resources. Depending on the resources required, some physiological measures may appear to be very sensitive in one task but not in another. For example, the duration and frequency of eye blinks decrease according to the quantity of visual information to be processed (Veltman & Gaillard, 1998), so this measure may appear irrelevant in a task that is mainly about memory and reasoning. Wickens and Hollands identified similar selectivity of the measure for the P300 amplitude of the event-related brain potential. They noted that the P300 amplitude decreases when a tracking task must be done concurrently with a primary task of counting tones. However, this amplitude was not sensitive to changes in the demands of the additional task, and the authors concluded that the two processes must have been drawing on different resources.

Because of this lack of selectivity and sensitivity, the relevance of using a physiological measure for mental workload assessment has to be validated in different tasks. It has been noted that, in doing this, some physiological techniques provide information relevant to a wide variety of tasks. For instance, the sensitivity of respiration rate to changes in task demands has been noted in a particular memory task (Backs & Seljos, 1994) as well as in air traffic control (Backs, Navidzadeh, & Xu, 2000). Furthermore, it is well known that mental workload can be measured using changes in pupil dilation (Kahneman & Beatty, 1966). For instance, during text reading, dilation increases when the reader encounters text of greater syntactic complexity (Just & Carpenter, 1993).

The pupil-dilation measurement technique is considered generally to be very sensitive[3] and to be a relevant approach for investigating many tasks that involve processing, including perception, memory, reasoning, and reading (Beatty, 1982; Peavler, 1974). Indeed, numerous studies have shown a strong correlation between pupil dilation and mental workload; however, the main inconsistency between them appears when task demands exceed participants' resources. For example, Granholm et al. (1996) noted that pupil dilation changes little near resource limits and starts to decline when demands exceed available resources. This is an interesting finding for identifying resource limits or resource allocations. Therefore, although this inconsistency deserves to be taken into account, the pupil-dilation measurement technique is considered to be very sensitive and to be one of the most relevant physiological measures of mental workload.

The Benefits of Multiple Mental Workload Measurements

In this brief description of three main categories of mental workload measurement, we selected a good representative within each category: the dual-task paradigm, the NASA-TLX rating scale, and the pupil-dilation measurement technique. For the combination of different methods to be justifiable, the minimal requirement is that no method perfectly measures the mental workload by itself. Each method has to provide a specific improvement when combined with another. To check this, we selected three components that are often the focus of research about mental workload: the requirements imposed by the techniques, the quantity of the data obtained, and the quality of the measure.

Requirement Imposed by the Techniques

To compare the requirements of the selected techniques, three criteria can be drawn from existing studies (e.g., Jex, 1988): the intrusiveness of the technique (Does the measure interfere with the main task?), the methodological requirements (What are the limitations of the measurement in capturing the load?), and the availability of the technique (Is the technique readily available?).

Intrusiveness of the technique. The three selected methods display low levels of intrusiveness. For the dual-task paradigm, the distribution of the signal is selected generally so as to ensure that it does not interfere with the primary task. The researcher must instruct the participant to focus on the main task rather than on the additional task, which could interfere with the primary task (Piolat et al., 1999). With regard to subjective measures, most researchers assume that intrusiveness is not a problem, particularly because data is collected after the task is complete. The measurement of pupil dilation does not require any action from the participant and so does not interfere with the main task (unless older eyetrackers are used that restrict the participants' movement). Therefore, intrusiveness alone is insufficient to distinguish among the techniques.

Requirements of the technique. When one compares techniques, it is important to consider the control required over the experimental task. A lack of this control could lead to bias in the data collection. For example, the dual-task paradigm involves measuring reaction times from signals (e.g., tones). This implicitly requires a controlled sound environment with limited noise. Despite this, the dual-task paradigm displays little known bias when compared with subjective measures. Most studies have probably been carried out in the area of subjective measures in which different biases are known. Some examples are given below.

The effect of delayed reporting on workload ratings. The quality of ratings is clearly affected after 48 h, and care must be exercised after 15 min (Moroney et al., 1995).

The effect of the participant's performance in the task. Failure in performing the task may lead to a higher perception of the mental workload than is necessary for success (Miyake, 2001).

The context effect from previous trials. If a task with moderate complexity is preceded by high load levels, the rating of the mental workload of the task tends to decrease, whereas, if a task with a similar level of complexity is preceded by lower load levels, the load rating tends to increase (Colle & Reid, 1998; Moroney et al., 1995).

The possibility of bias of this type calls for careful experimental design. For example, the pupil-dilation measurement technique makes specific demands, because luminosity is also known to be a factor determining pupil size. Consequently, pupil dilation should not be measured in field studies where the ambient light itself, or its effect on the measure, is not sufficiently controlled. Even relatively simple laboratory studies involving a computer task may require complex algorithms to distinguish the effects of changes in the brightness of the computer screen from those of mental workload on pupil dilation (e.g., Pomplun & Sunkara, 2003).

Availability of the technique. The NASA-TLX is the most readily available technique, because it requires only paper and pencil. The dual-task paradigm requires specific software to give the signals and collect the reaction times. However, as we describe below, one function of Tholos software is to handle the dual-task paradigm, so this technique can also be described as being available. Finally, the measurement of pupil dilation requires the most expensive apparatus due to the cost of most eyetracking systems, which makes this the least readily available technique.

These three criteria (intrusiveness, requirements, and availability) show that no single technique can be identified as the best. The subjective measures provide the cheapest way to determine mental workload, but they also display the highest level of known bias. The dual-task paradigm requires only software, but must be carried out in a very clean auditory (or visual) environment. Finally, the measurement of pupil dilation requires the most expensive system and the most control over the task environment.

The Quantity of Data in Mental Workload Assessment

The volume of data collected depends on the measurement technique used. As described in the following sections, some techniques yield a single index of the load, and others yield a more detailed picture of the dynamics of the load.

Dynamic versus static measurement. Subjective measurements are carried out after the task has been completed, whereas physiological or dual-task measures are carried out during the task. As a result, only the latter two can be used to explore the within-task variations in mental workload. More precisely, Xie and Salvendy (2000) considered several types of load, such as instantaneous load (the load at a specific moment), average load (the load per unit of time), accumulated load (the whole load experienced during the task), peak load (the maximum value of the instantaneous load), and overall load (which is determined a posteriori by subjective reports and is considered to result from both the average and the accumulated load). A subjective measure assesses only the overall load, whereas physiological and performance measures can be used to assess the other loads.

Low versus high bandwidth. Mental workload evolves in response to changes in the demands of the task, and so

in a dynamic approach, measurements must be taken frequently enough to detect transient changes (Wickens & Hollands, 2000). As a result, a physiological measure could miss fewer transient changes than would a dual-task paradigm (since the latter determines the load at less frequent intervals). Indeed, the intervals between the probes in the additional task have to be long enough not to interfere with the primary task. This means that the dual-task paradigm and the measure of pupil dilation differ in bandwidth.[4]

The pupil-dilation measurement technique has the most stringent requirements with regard to experimental settings (as we noted in the previous section), and, at the same time, it also offers the highest volume of data (it is a dynamic measure with a high bandwidth). The dual-task measurement has fewer requirements, but carries out only a partial measure of the load (it is a dynamic measure with a low bandwidth). Finally, a subjective measure, such as the NASA-TLX, has few requirements, but, because of its static nature, it offers the least data about the load. This means that we have to balance the requirements against the quantity of data to be collected. To complete these comparisons, it is also necessary to include the quality of data derived from the different measurements.

The Quality of Data in Mental Workload Assessment

Several authors have suggested criteria for evaluating the quality of the measurements of mental workload (Jex, 1988; Luximon & Goonetilleke, 2001; O'Donnell & Eggemeier, 1986; Wickens & Hollands, 2000). Three of these should be highlighted: the *sensitivity*, the *selectivity*, and the *diagnosticity* of the measures.

The sensitivity of the measure. The sensitivity level focuses on the ability of the method to discriminate among levels of task demands. Sensitivity is usually considered to be high for the pupil-dilation measurement technique; indeed, research has long suggested that pupil dilation correlates with mental workload. If one accepts the theoretical assumptions of the dual-task paradigm, its sensitivity may be considered as high in most cases if the signal frequency is sufficiently high. However, it may be considered to have low sensitivity if the primary task does not require many resources, because the performance of an additional task may be unaffected (Fisk et al., 1987). Subjective measures are considered to be very sensitive for comparing tasks (or trials), but may not be sensitive enough to assess the dynamics of the load within the same task.

The selectivity of the measure. In contrast, selectivity requires that the measure stay unchanged when the task load does not change. This is a problem particularly with respect to physiological measures. For example, pupil-dilation measurement technique presents a low level of selectivity because there is also noise due to external factors, such as changes in the ambient light. Lack of selectivity also results from individual factors that are more difficult to control. For example, the validity of pupil dilation for measuring emotional reactions has been clearly established (Kahneman, Peavler, & Onuska, 1968; Partala & Surakka, 2003). This explains why the mental workload should not always be invoked as the sole explanation of changes in pupil dilation. On the other hand, the subjective measurement, due to its bias (such as the effect of prior trials or the performance of the task), may be considered as not always selective. Therefore, the dual-task paradigm is probably the most selective of the three measurements.

The diagnosticity of the measure. The diagnosticity criterion relates to the capability of the measure to identify the source of the load in the task. For example, pupil-dilation measurement technique and the dual-task paradigm both have a low level of diagnosticity because they offer a global metric, regardless of the source of the load in the task. However, the multidimensional nature of NASA-TLX makes it the most relevant way of determining the source of the load. For instance, it can be used to identify whether overloading during the task results from a high level in the cognitive-demand dimension or in the temporal-demand dimension. The ability to make this distinction could provide important information in most field studies or complex laboratory experiments.

Toward Combining Several Methods

On the basis of the criteria listed above, no single technique emerged as the best.

The dual-task paradigm is capable of modeling the dynamics of the load, has low bandwidth, and is very selective. However, this method is not always sensitive and has zero diagnosticity.

The NASA-TLX rating scale has the fewest technical requirements and the lowest levels of intrusiveness, but has several methodological limitations. Moreover, it is a static measurement that cannot assess the dynamics of the load and is not sensitive to within-task changes in the load (nor is it very selective). However, it is the only measure of the three selected to display a high diagnosticity.

The pupil-dilation measurement technique requires a very expensive system and the most controlled environment. However, its high bandwidth provides the highest sensitivity for detecting the dynamics of the load. It must be noted, though, that this technique is not very selective and does not include any diagnosticity.

Table 1 shows a comparison of the three techniques based on the requirements, the quantity of the data, and the quality of the measure.

In the light of these considerations, we suggest that mental workload should be determined by combining several methods in order to compensate for these different limitations, as outlined in the following sections.

Combining measurements may increase sensitivity. A single measurement may not detect all the dynamics of the load. This is clearly the case for a low-bandwidth measure such as the dual-task paradigm. This is also true of a physiological measurement such as pupil dilation, because during eye blinks, a pupil diameter of zero is usually recorded (even for the duration of a short blink). At the same time, eye blink duration and intervals tend to change depending on the visual demands of the task (Veltman & Gaillard, 1998). This means that some very interesting pupil-dilation data could be lost during eye blinks. Consequently, combining the measurement of eye blinks with pupil dilation may increase the sensitivity. Adding a

Table 1
A Comparison of the Three Techniques on the Basis of Requirements of the Techniques, the Quantity of the Data, and the Quality of the Measure

	Dual-Task Paradigm	NASA-TLX	Pupil Dilation
Requirements			
Intrusiveness	Low	Low	Low (in laboratory)
Availability	High (using existing software)	High	Low
Requirements	Low	High	High
Quantity of Data			
Static or dynamic	Dynamic	Static	Dynamic
Bandwidth	Low	n/a	High
Quality of the Measure			
Sensitivity	High/low (depending on the demands of the primary task)	High (between-tasks)/ low (within-tasks)	High
Selectivity	High	Low	Low
Diagnosticity	Low	High	Low

method from another category, such as the dual-task paradigm, may increase sensitivity still further.

Combining measurements may increase selectivity. Some factors, particularly in physiological measures, are difficult to control—for example, brightness of the computer screen or ambient light. To increase the selectivity of the measurement, it may therefore be useful to add another measure. For example, Veltman and Gaillard (1998) noted that variability in heart rate and blood pressure provides sensitive measurement of the load, whereas respiration rate is less sensitive. They also noted that respiration rate affects both heart rate and blood pressure. They suggested, therefore, either measuring the respiration rate to correct one of the other measurements or combining heart rate and blood pressure measurements to form a new index intended to minimize interference from the respiration rate. Either way, selectivity is increased.

Combining measurements may increase diagnosticity. As noted, the subjective reports are sensitive to between-task differences, but not to within-task ones. Therefore, one may be inclined to exclude these measures when dynamic measurements are available. However, depending on the dimension of the scale, subjective reports may provide useful complementary data by evaluating the nature of the demands experienced by the participant.

Combining multiple measurements of mental workload may increase the quality and the quantity of available data. However, the technical requirements for combining measures are very high, which may explain why relatively few researchers have attempted to compare methods from different categories (but see, e.g., Bortolussi, Kantowitz, & Hart, 1986; Casali & Wierwille, 1983; Isreal, Chesney, Wickens, & Donchin, 1980; James et al., 1995; Miyake, 2001). It was in order to reduce these technical requirements that we designed a specific software package, named Tholos.

THOLOS SOFTWARE
A Tool for Combining Measurements of Mental Workload

Tholos software was designed to facilitate the following: (1) the design of experimental studies that measure mental workload (Tholos provides easily customizable performance and subjective measurement techniques of mental workload, thus reducing the technical requirements of software development); (2) the collection of data (Tholos produces standard spreadsheet documents that can be imported, for example, into OpenOffice Calc or Microsoft Excel); (3) data analysis (Tholos can combine performance and physiological measures into a single protocol). These functions all contribute to achieving the main goal of Tholos: to help experimenters combine various measures of mental workload.

Measures Implemented Using Tholos

Tholos and performance measures. Tholos implements several methods of measuring mental workload. From the available performance measure methods, we have included the dual-task paradigm with auditory signals in the software. As noted, to limit its intrusiveness, the dual-task paradigm requires careful examination of the disruption to the primary task: The distribution of auditory signals has to be arranged so as to prevent the participants from focusing on the additional task. Piolat et al. (1999) also stressed that determination of signal distribution must be based on the research objectives and on the characteristics of the primary task. So when using Tholos, the experimenter must specify a range of distribution for the auditory signal (see Figure 1) and not a fixed interval. The latter prevents the predictability of the auditory signal, which may lead to automatization (Fisk et al., 1987).

In order to control individual differences in reaction times, Tholos requires participants to undertake a training session before performing the task (and responding to auditory signals). This session makes it possible to establish a base value for the participant's reaction time. This value is later subtracted from the reaction time recorded during the experiment. This increases the selectivity of the measure.

During the task, there are auditory signals to which the participant must react with a mouse click or a keypress. The higher the reaction time, the more cognitive resources are considered to be involved in the primary task. At the end of the task, the Tholos software generates a spreadsheet file. This can be opened with most spreadsheet soft-

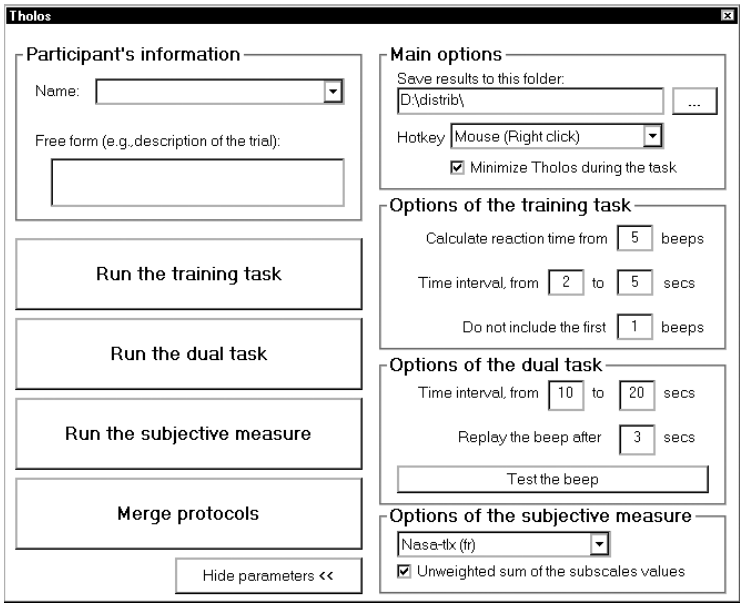

Figure 1. The Tholos main screen. The top left section is used for information about the participant; the right section is used to configure the performance and the subjective measures; the bottom left section is used to start the training, the dual task, and the rating scales.

ware and contains several values, such as auditory signal distribution and the participants' reaction times. Consequently, with Tholos, it is possible to study the dynamics of the mental workload while the participant is performing the experimental task (the primary task).

Tholos and subjective measures. The Tholos software also implements the NASA-TLX rating scales. By using the main Tholos interface, the experimenter may choose between the simplified RTLX version of the calculation and the traditional version with pairwise comparisons of the scales (see the "Subjective measures" section above). It is also possible to modify the content of the scales in order to translate them, or to promote a variant of the scales in order to adapt them to the specificities of the studied task. For example, one could change the content of the scales for specific tasks, such as car driving (e.g., DooWon & Peom, 1997). After the main task has been completed, the experimenter can click on a button to display the NASA-TLX. The rating scales appear on the screen, and the participant can set the rating for each question by using a slider (see Figure 2).

If the experimenter selects the traditional version of the scale, the second part starts, during which the participant has to weight each of the dimensions in pairs, resulting in 15 comparisons. Finally, as for the dual-task paradigm, the results of the ratings and of the overall load are recorded in a spreadsheet for facilitating analyses.

Tholos and physiological measures. The goal of the Tholos software is not to replace existing tools designed for physiological measurement (especially for the measure of pupil dilation), but to associate their data with other measurements. To this end, Tholos merges data collected from physiological tools with data from the dual-task paradigm (i.e., reaction times). The experimenter can select data from physiological measures and performance measures (dual task) to generate a new spreadsheet document. This new document combines all the data along the time dimension, which (as the next section shows) makes it possible to compare the measurements of mental workload.

Studies involving pupil-dilation measurement face the problem of eye blinks, during which a pupil diameter of zero is usually recorded. At the same time, one may observe unusual values due to the partial obscuration of the pupil. By activating the corresponding option in the Tholos interface, the experimenter can remove blinks and partial blink artifacts (see Figure 3). This is done by removing all zero values, as well as extreme values within 100 msec (a blink generally lasts for 70–100 msec), as suggested by Shi et al. (2003). Moreover, new formats relating to other physiological measures can be added easily in Tholos, allowing the experimenter to combine the dynamics of the load from a large variety of sources.

A Case Study of Puzzle Solving

To offer a simple example of how Tholos works, we designed an experiment based on the sudoku puzzle. The aim of this puzzle is to fill a 9×9 grid with numerical digits so that every row, every column, and every 3×3 box contains the digits from 1 to 9. The puzzle may be defined as a constraint-satisfaction problem that requires the participant to formulate, propagate, and satisfy a large number of constraints. In order to fill an empty cell in the grid, a participant must identify the cells that impose constraints on the cell under consideration. By putting a value in this cell, the participant also adds constraints, which may allow new actions. The number of new actions

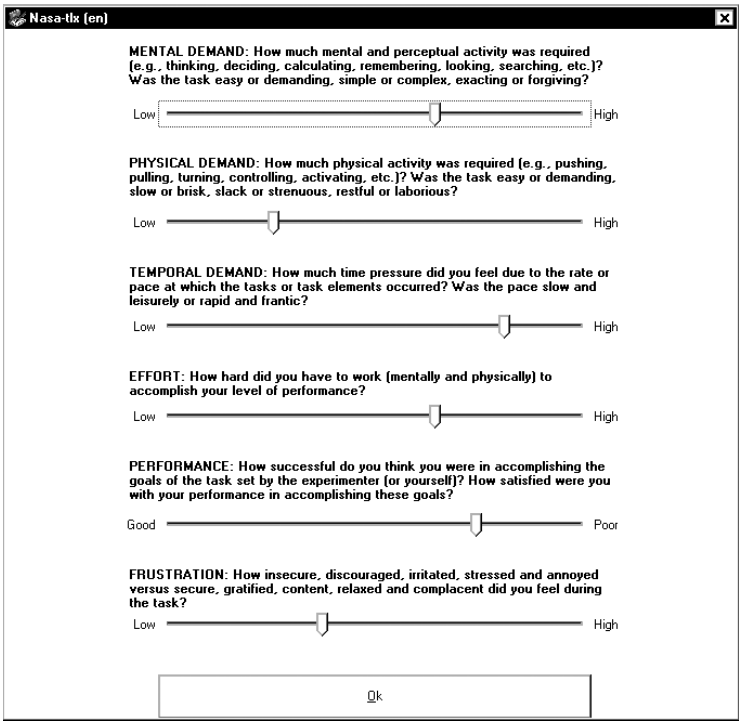

Figure 2. Tholos displaying the NASA-TLX scales; the participants have to rate each dimension.

available usually changes, depending on the complexity of the puzzle. In a relatively easy puzzle, a large number of remaining cells may be solved independently of other cells. In a more complex puzzle, the participant must identify specific constraints in order to access a few others, which, in turn, allows other cells to be solved; that is, there are several "locks." In this experiment, 4 participants who were familiar with sudoku puzzles (having already solved at least 20 such puzzles) had to solve two puzzles with two levels of complexity (low and high). The low-complexity puzzle started with 45 empty cells and three locks to be solved; the high-complexity puzzle started with 50 empty cells and six locks.

Materials. During puzzle solving, we successively recorded 4 participants' pupil dilation using the SMI iViewX head-mounted eyetracking system, which has a 50-Hz

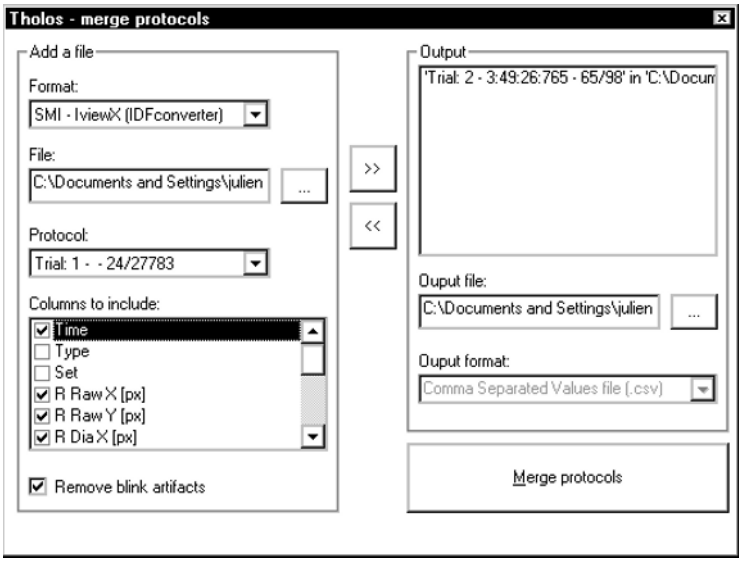

Figure 3. Using the Tholos interface to merge protocols (in this case, from an eyetracking protocol and from the dual-task protocol).

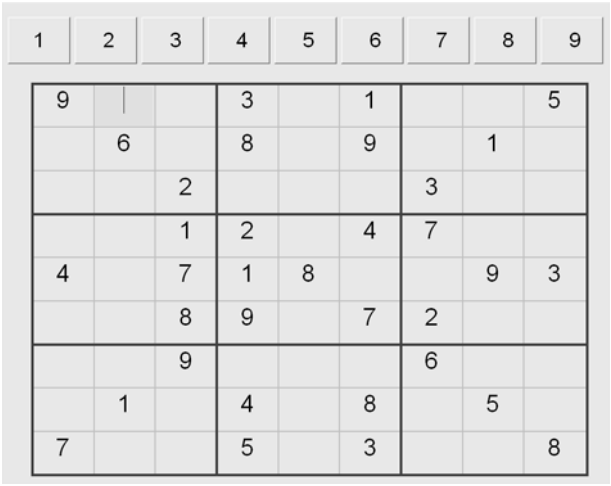

Figure 4. The experimental task based on the sudoku puzzle. Each participant has to fill the grid, selecting the numerical digits so that every row, every column, and every 3 × 3 box contains the digits from 1 to 9. The colors of the figure have been modified for publication.

sampling rate. The ambient light was controlled during the experiment, which took place in a noise-attenuated room. While the participants performed the task, we used Tholos to record reaction times to standard tones played through the computer's speakers at 10- to 20-sec intervals. The intrusiveness of the signal range was controlled in a pilot study. After each puzzle, we used Tholos to assess the subjective rating of the mental workload with the NASA-TLX. When the experiment was over, we imported the participants' pupil diameter into Tholos to combine data from the eyetracker with the dual-task reaction times. This made it possible to obtain three measures of mental workload and thus to increase the variety of the results.

Procedure. The experimental task was divided into the following stages: Participants were trained to respond to the auditory signals as rapidly as possible so that the base value of the reaction time could be calculated; participants were given the computerized version of the puzzle so they could become familiar with the task and with the experimental software (see Figure 4); and participants had to solve the two puzzles. Half of the participants were tested on the less complex puzzle first; the other half were tested in the reverse order.

Results.[5] The easy puzzle was solved in about 16 min versus about 45 min for the more complex one. The unweighted sum of the NASA-TLX scales indicates a score of 39.37 for the less complex puzzle versus 48.1 for the more complex one. These scores suggest that the sensitivity of the rating scales made it possible to distinguish between the two levels of complexity. Moreover, because the rating is multidimensional, it is possible to identify the source of the increase in mental workload. More precisely, the values of the dimensions indicate three main differences according to the complexity level (see Figure 5). There was an increase in the mental demand, effort, and frustration dimensions, whereas the other dimensions remained almost stable (a difference of 1% or less for physical demand, temporal demand, and performance was considered to be negligible). These results indicate that, in the more complex puzzle, participants experienced an increase in the quantity of cognitive activities and were less confident but did not experience any more physical or temporal demand. For puzzles of either complexity level, participants were satisfied with their performance, as indicated by the performance dimension (a score of 0 indicates a good performance; a score of 100 indicates a poor performance).

For the dynamic measure (reaction time and pupil dilation), we calculated the accumulated load (i.e., the sum of all instantaneous load measurements), peak load (i.e., the highest instantaneous load), and average load (i.e., the accumulated load divided by the number of instantaneous load measurements). These are shown in Table 2, which also shows the overall load (i.e., RTLX).

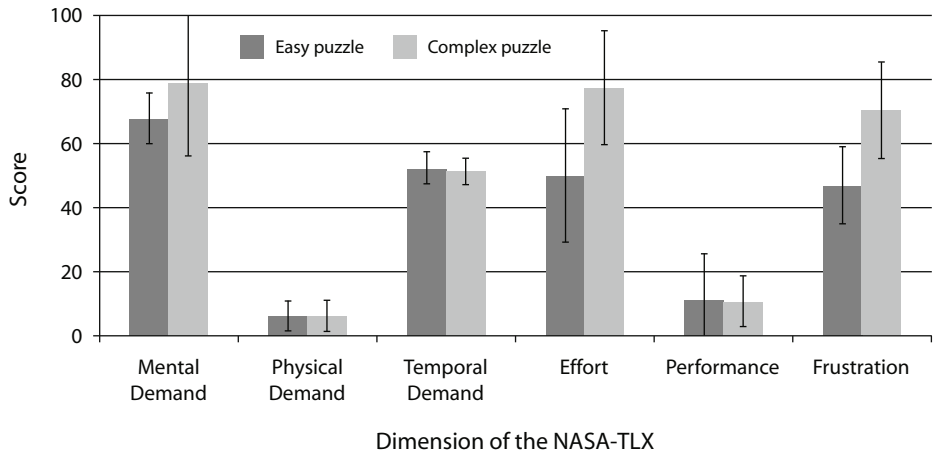

Figure 5. NASA-TLX results depending on the complexity level of the puzzles. The higher the bar, the higher the load on this dimension, except for the "performance" dimension, where the lower the value the higher (or better) the performance. When comparing the two complexity levels, it is possible to assess the main sources of the overall load.

Table 2
Measures of Mental Workload for Dynamic (Dual-Task Paradigm and Pupil Dilation) and Static (NASA-TLX) Measures: Average Load (the Load per Unit of Time), Accumulated Load (the Whole Load Experienced During the Task), Peak Load (the Maximum Value of Instantaneous Load), and Overall Load (Subjective Reports)

	Dual-Task Paradigm (msec)				Pupil Dilation (mm)				NASA-TLX (Rating Scale From 0 to 100)			
	Easy		Complex		Easy		Complex		Easy		Complex	
Complexity Level	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Average load	170.52	38.37	277	60.03	3.4	0.624	3.54	0.31				
Accumulated load	6,997.76	912.64	24,220.96	8,732.75	137,807.34	40,318.7	388,903.07	161,109.72				
Peak load	483.55	58.97	1,250.61	571.25	4.16	0.44	4.25	0.42				
Overall load									39.37	5.1	48.1	5.9

On the basis of the reaction time (i.e., the dual-task paradigm), the accumulated load was 3.4 times greater for the complex puzzle than for the easy puzzle. The peak load was about 2.6 times greater for the complex puzzle. Finally, the average load was greater for the complex puzzle (277 msec) than for the easy one (170.52 msec). These results show the dual-task paradigm's ability to discriminate changes in the load under the two conditions, as is also indicated in the subjective reports using the NASA-TLX rating scales.

The pupil-dilation results were quite similar. The accumulated load was 2.8 times higher for the most complex scenario, which is consistent with the duration of the scenario. The peak load was higher for the complex puzzle than for the easy one (4.25 vs. 4.16 mm). A smaller difference was also found for the average load (3.54 vs. 3.4 mm), and such a difference was considered to be significant according to Ahlstrom and Friedman-Berg (2006). These results stress the sensitivity of these measures to differences in mental workload depending on the complexity of the puzzle. To consolidate this claim, we suggest that a more detailed analysis, participant by participant, should compare the dynamics of these last two measures.

As indicated previously, the Tholos software facilitates such a comparison. Figure 6 illustrates a combination for the easiest of the two puzzles for 1 participant (other participants display similar results). Three parameters are plotted along a time axis. The reaction time was measured at 10- to 20-sec intervals, shown on a scale from 0–400 msec, and displayed as a curve (rather than as discrete plots) in order to facilitate comparison. Pupil diameter is shown on a scale from 3 to 4.5 mm (we fitted a smoothing spline to this data set, which reduces the noise of the measure). A data comparison is made between the reaction time curve and the (smoothed) pupil diameter curve. The participant's puzzle-solving actions (entering new values) are plotted with asterisks according to their time occurrence.

Indeed, reaction time and pupil diameter cannot be compared on the basis of exact values, because different unit measurements (millimeters and milliseconds) are used for the scales. However, the shapes of the graphs provide relevant information. By looking at both graphs, we can see that they follow a very similar pattern. For example, until peaking after about 3 min, they display successive increases and decreases, which relate to puzzle-solving activities. However, on closer examination of the two graphs, differences in shape become apparent.

For example, the changes do not occur at exactly the same time. The peaks in the pupil-dilation graph sometimes precede the corresponding peak in the reaction-time graph (e.g., at about 3 min on the *x*-axis), sometimes seem to occur just after this peak (e.g., at about 5 min), and sometimes seem to be unrelated to the other graph (e.g., at about 7 min).

Also, the graphs follow different scales. As noted previously, the units on the two graphs differ, which means that the extent of the rise or fall of the graph also differs. For example, a small increase or decrease in pupil diameter may be reflected by much larger variations in the reaction times (e.g., at about 5 min).

Also, the physiological measure detects more changes. This may be easily explained by the sampling difference. Looking at the smoothed graph of pupil dilation, we can

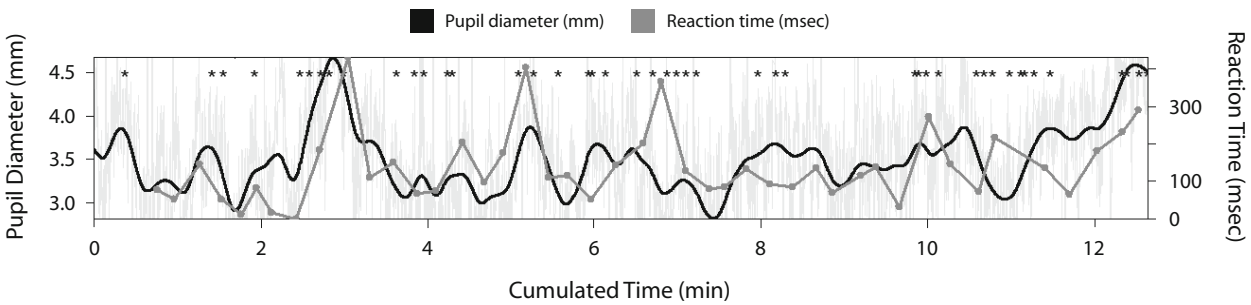

Figure 6. Sample of one participant's graph of the reaction time (in milliseconds) and pupil-dilation (in millimeters) measures versus time for the less complex puzzle. The asterisks indicate the moment when the participant fills an empty cell of the puzzle.

see that the reaction time graph often badly approximates its shape. This implies that sensitivity of the curve is indeed dependent on the sampling of the measure.

From a methodological point of view, the combination of different measures provides useful information for the experimenter. In fact, it is almost impossible to discuss the changes in mental workload for the reaction times along the time axis without using a second measure. For example, we must be careful about claiming that an increase in load is followed by a decrease (or vice versa), because the shape of the curve between the known values is uncertain. In particular, when sampling frequency is low, it is possible to miss several peaks in the load. There are also several cases in which peaks in the reaction times were not accompanied by changes in the pupil-dilation graph. Because pupil dilation is an indication of puzzle-solving activity, its absence may have resulted partly from missing values due to blinking.

Another explanation may lead to a discussion of the diagnosticity of the pupil-dilation measurement technique. As noted by Wilson and O'Donnell (1988), an increased load does not always result in increased overall activation. The more selective measure would detect changes in the load, whereas the general measure of load would not. This would also mean that, even if the changes in the two graphs seem to imply a close correlation, one must interpret the results carefully, because the changes do not necessarily have a common cause.

Indeed, although these are exploratory results, they illustrate how Tholos can facilitate research by supporting the combination of multiple measures; although the diagnosticity of the NASA-TLX improves understanding of the source of the load, it does not reflect the dynamics of the load. The dual-task paradigm appears to be a good candidate method for comparing loads in different trials, whereas the pupil-dilation measurement technique appears to be a good way to measure load changes within each trial. However, such claims require support from further experimental studies. Various theoretical and methodological questions arise from combining measures, as the Tholos software does. We discuss such questions related to the puzzle-solving case study in the following section.

DISCUSSION

As we indicated above, combining several measures of mental workload may contribute to improving both the methodological and theoretical foundations of mental workload. For example, several methodological questions arise from these experimental findings, and the answers to these questions may clarify the relevance and quality of the different techniques.

The first question concerns those situations for which the dual-task paradigm is not appropriate. The distribution intervals of the probes have to be identified in a pilot study to limit the intrusiveness of the technique (Piolat et al., 1999). However, in specific situations, a very long interval may be required, and this may lead to many transient changes being missed. In an extreme situation, missing too many changes could lead to an irrelevant measure that incorrectly reflects the changes in the task load. Moreover, because the graphs reveal large differences between reaction times and pupil dilation, the measures must be validated more precisely. Future research must address this point by comparing more precisely the correlation between the changes in the dual-task measure and changes in the measure of pupil dilation in different tasks with specific distribution intervals.

Another methodological question concerns the links between the overall load measured by a static measure (such as the NASA-TLX) and that measured dynamically. Xie and Salvendy (2000) demonstrated that the overall load is different from both the accumulated load and the average load, even though they share some relations. Miyake (2001) and Ryu and Myung (2005) suggested that it is possible to gain an accurate understanding of participants' reports by analyzing dynamic (physiological) measures of the mental workload. To obtain a better understanding of these links, one might consider assessing the various dimensions of the NASA-TLX separately by using different measures. For example, using a wide variety of physiological measures may provide a relevant method of assessing the individual dimensions of the scales separately.

This last question also implies that theoretical improvements of the mental workload concept could be obtained as a result of a more precise understanding of the components that are under consideration. From a theoretical point of view, another question arises from the results of the dual-task paradigm. As stated by Fisk et al. (1987), a secondary task that is sensitive to changes in the demands of the primary task should be selected from within an identical pool of resources. This consideration refers to the multiple-resource theory of Wickens (1984). However, Piolat et al. (1999) suggested that a probe should be selected in another modality, referring here to the single-resource theory of Kahneman (1973). The claim that the reaction time resulting from a probe in another modality is an index of mental workload requires careful scrutiny. A more detailed comparison of physiological measures and the dual-task paradigm, as suggested by Isreal et al. (1980), might provide more details about the theoretical foundations of the dual-task paradigm and about the validity of selecting a probe in another modality. This may take the form of a comparison between the dual-task paradigm and the irrelevant-probe technique (Papanicolaou & Johnstone, 1984). This technique involves recording of physiological measures (such as event-related potentials), in which, unlike in the dual-task paradigm, the participant has to ignore the probes. Several authors have noted that attention not focused elsewhere is attracted to the probes, even if the participant does not pay attention directly to them (Kramer, Trejo, & Humphrey, 1995; Ullsperger, Freude, & Erdmann, 2001).

Three questions have emerged from combining multiple measures of mental workload. These questions may help us to build upon existing theoretical and methodological foundations of mental workload. They also demonstrate the benefits of Tholos software in reducing the requirements of comparing several measures of mental workload.

CONCLUSION

No single measure, be it performance, subjective, or physiological, can provide a better overall assessment of mental workload than any other. We have therefore suggested combining several measures in order to increase the sensitivity, selectivity, and diagnosticity of the mental workload determination. We developed Tholos software to reduce the implementation requirements. Tholos facilitates the experimenter's job in collecting data from subjective reports and from the dual-task paradigm, while making it possible to combine several measures of the dynamics of the mental workload.

Moreover, Tholos can be used in various types of experiment, such as driving a simulator, controlling air traffic, or using a computer interface. In our experiment, combining the different types of measures highlighted a number of theoretical and methodological issues. This illustrates the usefulness of Tholos in helping experimenters reach a clearer understanding of the nature of the mental workload concept. We hope that this will promote the widespread acceptance of the measures and will lead toward a more consensual definition of this notion.

AUTHOR NOTE

We are very grateful to Monika Gosh for proofreading this article. We also thank the anonymous reviewers for their helpful comments. Address correspondence to J. Cegarra, Université de Toulouse, CLLE-LTC, Centre Universitaire Champollion, Place de Verdun, 81012 Albi cedex 9, France (e-mail: julien.cegarra@univ-jfc.fr).

REFERENCES

Ahlstrom, U., & Friedman-Berg, F. J. (2006). Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics*, **36**, 623-636.

Backs, R. W., Navidzadeh, H. T., & Xu, X. (2000). Cardiorespiratory indices of mental workload during simulated air traffic control. *Proceedings of the Human Factors and Ergonomics Society 44th Annual Meeting* (pp. 3-89). Santa Monica, CA: Human Factors and Ergonomics Society.

Backs, R. W., & Seljos, K. A. (1994). Metabolic and cardiorespiratory measures of mental effort: The effects of level of difficulty in a working memory task. *International Journal of Psychophysiology*, **16**, 57-68.

Bainbridge, L. (1974). Problems in the assessment of mental load. *Le Travail Humain*, **37**, 279-302.

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

Bi, S., & Salvendy, G. (1994). Analytical modeling and experimental study of human workload in scheduling of advanced manufacturing systems. *International Journal of Human Factors in Manufacturing*, **4**, 205-234.

Bortolussi, M. R., Kantowitz, B. H., & Hart, S. G. (1986). Measuring pilot workload in a motion base trainer: A comparison of four techniques. *Applied Ergonomics*, **17**, 278-283.

Brünken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, **38**, 53-61.

Byers, J. C., Bittner, A. C., Jr., & Hill, S. G. (1989). Traditional and raw task load index (TLX) correlations: Are paired comparisons necessary? In A. Mital (Ed.), *Advances in industrial ergonomics and safety* (pp. 481-485). London: Taylor & Francis.

Casali, J. G., & Wierwille, W. W. (1983). A comparison of rating scale, secondary task, physiological, and primary-task workload estimation techniques in a simulated flight task emphasizing communications load. *Human Factors*, **25**, 623-641.

Cegarra, J., & Hoc, J.-M. (2006). Cognitive styles as an explanation of experts' individual differences: A case study in computer-assisted troubleshooting diagnosis. *International Journal of Human–Computer Studies*, **64**, 123-136.

Chevalier, A., & Kicka, M. (2006). Web designers and Web users: Influence of the ergonomic quality of the Web site on the information search. *International Journal of Human–Computer Studies*, **64**, 1031-1048.

Colle, H. A., & Reid, G. B. (1998). Context effects in subjective mental workload ratings. *Human Factors*, **40**, 591-600.

Cooper, G. E., & Harper, R. P., Jr. (1969). *The use of pilot rating in the evaluation of aircraft handling qualities* (Tech. Rep. D-5153). Washington, DC: NASA.

DooWon, C., & Peom, P. (1997). Simulator-based cognitive load assessment of the in-vehicle navigation system driver using revision of NASA-TLX. *IE-Interfaces*, **10**, 145-154.

Fisk, A. D., Derrick, W. L., & Schneider, W. (1987). A methodological assessment and evaluation of dual-task paradigms. *Current Psychological Research & Reviews*, **5**, 315-327.

Granholm, E., Asarnow, R. F., Sarkin, A. J., & Dykes, K. L. (1996). Pupillary responses index cognitive resource limitations. *Psychophysiology*, **33**, 457-461.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139-183). Amsterdam: North-Holland.

Hendy, K. C., Hamilton, K. M., & Landry, L. N. (1993). Measuring subjective workload: When is one scale better than many? *Human Factors*, **35**, 579-601.

Hill, S. G., Iavecchia, H. P., Byers, J. C., Bittner, A. C., Zaklad, A. L., & Christ, R. E. (1992). Comparison of four subjective workload rating scales. *Human Factors*, **34**, 429-439.

Isreal, J. B., Chesney, G. L., Wickens, C. D., & Donchin, E. (1980). P300 and tracking difficulty: Evidence for multiple resources in dual-task performance. *Psychophysiology*, **17**, 259-273.

James, R. H., Elderfield, H., Palmer, M. R., & Connelly, C. S. (1995). Toward an understanding of DCS control operator workload. *ISA Transactions*, **34**, 175-184.

Jex, H. R. (1988). Measuring mental workload: Problems, progress, and promises. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 5-39). Amsterdam: North-Holland.

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., Peavler, W. S., & Onuska, L. (1968). Effects of verbalization and incentive on the pupil response to mental activity. *Canadian Journal of Psychology*, **22**, 186-196.

Kellogg, R. T. (1990). Effectiveness of prewriting strategies as a function of task demands. *American Journal of Psychology*, **103**, 327-342.

Kilmer, K. J., Knapp, R., Burdsal, C., Borresen, R., Bateman, R., & Malzahn, D. (1988). Techniques of subjective assessment: A comparison of the SWAT and modified Cooper–Harper scales. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 155-159). Santa Monica, CA: Human Factors Society.

Kramer, A. F., Trejo, L. J., & Humphrey, D. (1995). Assessment of mental workload with task-irrelevant auditory probes. *Biological Psychology*, **40**, 83-100.

Lee, D. H., & Park, K. S. (1990). Multivariate analysis of mental and physical load components in sinus arrhythmia scores. *Ergonomics*, **33**, 35-47.

Levy, C. M., & Ransdell, S. (1995). Is writing as difficult as it seems? *Memory & Cognition*, **23**, 767-779.

Luximon, A., & Goonetilleke, R. S. (2001). Simplified subjective workload assessment technique. *Ergonomics*, **44**, 229-243.

Lysaght, R. J., Hill, S. G., Dick, A. O., Plamondon, B. D., Linton, P. M.,

WIERWILLE, W. W., ET AL. (1989). *Operator workload: Comprehensive review and evaluation of operator workload methodologies* (Tech. Rep. 851). Fort Bliss, TX: U.S. Army Research Institute, Field Unit.

MIYAKE, S. (2001). Multivariate workload evaluation combining physiological and subjective measures. *International Journal of Psychophysiology*, **40**, 233-238.

MORONEY, W. F., BIERS, D. W., & EGGEMEIER, F. T. (1995). Some measurement and methodological considerations in the application of subjective workload measurement techniques. *International Journal of Aviation Psychology*, **5**, 87-106.

NORMAN, D. A., & BOBROW, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, **7**, 44-64.

O'DONNELL, R. D., & EGGEMEIER, F. T. (1986). Workload assessment methodology. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance* (Vol. 2, pp. 1-49). New York: Wiley.

PAAS, F., TUOVINEN, J. E., VAN MERRIËNBOER, J. J. G., & DARABI, A. A. (2005). A motivational perspective on the relation between mental effort and performance: Optimizing learner involvement in instruction. *Journal of Educational Technology, Research, & Development*, **53**, 5-11.

PAPANICOLAOU, A. C., & JOHNSTONE, J. (1984). Probe evoked potentials: Theory, method and applications. *International Journal of Neuroscience*, **24**, 107-131.

PARTALA, T., & SURAKKA, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human–Computer Studies*, **59**, 185-198.

PEAVLER, W. S. (1974). Pupil size, information overload, and performance differences. *Psychophysiology*, **11**, 559-566.

PIOLAT, A., OLIVE, T., ROUSSEY, J.-Y., THUNIN, O., & ZIEGLER, J. C. (1999). SCRIPTKELL: A tool for measuring cognitive effort and time processing in writing and other complex cognitive activities. *Behavior Research Methods, Instruments, & Computers*, **31**, 113-121.

POMPLUN, M., & SUNKARA, S. (2003). Pupil dilation as an indicator of cognitive workload in human–computer interaction. *Proceedings of HCI International 2003* (Vol. 3, pp. 542-546). Mahwah, NJ: Erlbaum.

REID, G. B., POTTER, S. S., & BRESSLER, J. R. (1987). *User's guide for the Subjective Workload Assessment Technique (SWAT)* (Tech. Rep. AAMRL-TR-87-023), Wright-Patterson AFB, OH: Armstrong Aerospace Medical Research Laboratory.

RUBIO, S., DÍAZ, E., MARTÍN, J., & PUENTE, J. M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied Psychology*, **53**, 61-86.

RYU, K., & MYUNG, R. (2005). Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics*, **35**, 991-1009.

SCHVANEVELDT, R. W., REID, G. B., GOMEZ, R. L., & RICE, S. (1998). Modeling mental workload. *Cognitive Technology*, **3**, 19-31.

SHI, B., MOLONEY, K., EMERY, V., JACKO, J., SAINFORT, F., & VIDAKOVIC, B. (2003). *Multifractal discrimination model of high-frequency pupil-diameter measurements.* Working paper.

ULLSPERGER, P., FREUDE, G., & ERDMANN, U. (2001). Auditory probe sensitivity to mental workload changes—An event-related potential study. *International Journal of Psychophysiology*, **40**, 201-209.

VELTMAN, J. A., & GAILLARD, A. W. K. (1998). Physiological indices of workload in a simulated flight task. *Biological Psychology*, **42**, 323-342.

WICKENS, C. D. (1984). Processing resources in attention. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention* (pp. 63-102). San Diego, CA: Academic Press.

WICKENS, C. D., & HOLLANDS, J. (2000). *Engineering psychology and human performance*. London: Longmans.

WIERWILLE, W. W., & CASALI, J. G. (1983). A validated rating scale for global mental workload measurement applications. In *Proceedings of the 27th Annual Meeting of the Human Factors Society* (pp. 129-133). Santa Monica, CA: Human Factors and Ergonomics Society.

WILSON, G. F., & O'DONNELL, R. D. (1988). Measurement of operator workload with the neuropsychological workload test battery. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 63-100). Amsterdam: Elsevier.

XIE, B., & SALVENDY, G. (2000). Prediction of mental workload in single and multiple tasks environments. *International Journal of Cognitive Ergonomics*, **4**, 213-242.

NOTES

1. In ancient Greek architecture, a *tholos* was a dome-shaped building in which officials kept weights and measures. Tholos software can be downloaded from tholos.psychologie-fr.org.

2. The reader can refer to Piolat, Olive, Roussey, Thunin, and Ziegler (1999) for a theoretical discussion about the dual- (and triple-) task paradigms. The authors have developed a specific tool, known as SCRIPTKELL, which also supports the triple-task paradigm.

3. Selectivity of the pupil-dilation measurement technique is discussed in the section "The Quality of Data in Mental Workload Assessment."

4. Static measures, such as subjective reports, are excluded from this dimension because they are carried out after the task has been completed. Therefore, they should not be confused with a very low-bandwidth measure.

5. Because the goal of this article is not to analyze puzzle-solving cognitive processes precisely but only to illustrate Tholos's usefulness in combining measures, we provide simple summaries of the results based on descriptive statistics.

(Manuscript received December 16, 2007;
revision accepted for publication March 13, 2008.)