Verbalization rate as an index of cognitive load

VERBALIZATION RATE AS AN INDEX OF COGNITIVE LOAD

by

J.A. Brinkman

Report EUT/BDK/58
ISBN 90-386-0351-7
ISSN 0929-8479
Eindhoven 1993

Eindhoven University of Technology
Graduate School of Industrial Engineering
and Management Science
Eindhoven, The Netherlands

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Brinkman, J.A.

Verbalization rate as an index of cognitive load / by J.A. Brinkman.
Eindhoven: Eindhoven University of Technology,
Graduate School of Industrial Engineering and Management Science.
Report EUT/BDK, ISSN 0929-8479;58.
With. ref.
ISBN 90-386-0351-7
Subject headings: verbal protocol analysis.

VERBALIZATION RATE AS AN INDEX OF COGNITIVE LOAD[1]

J.A. Brinkman

Abstract

The study reported about dealt with the applicability of thinking aloud as a secondary task for measuring the cognitive load associated with a primary task. More specifically, it examined whether think-aloud performance, as indexed by the rate of verbalization, was sensitive to variations in primary-task difficulty. With this goal in mind, a group of 24 subjects had to verbalize their thoughts while carrying out a primary task the difficulty of which was experimentally manipulated. As expected, an increase in task difficulty led to a decrease in verbalization rate, at least when controlling for the way the task was actually carried out (i.e., strategy use). Unfortunately, straightforward interpretation of this decrease as representing the cognitive load of the primary task was not possible since thinking aloud altered primary-task performance significantly. Therefore, the results of the study do not warrant the conclusion yet that verbalization rate can be used as a sensitive index of cognitive load.

[1] The study reported here essentially involves a re-analysis of the data of an earlier experiment. This experiment was part of a larger study on the applicability of the verbal protocol methodology in the task domain of fault diagnosis. It is described fully in the author's Ph.D. thesis "The analysis of fault diagnosis tasks: do verbal reports speak for themselves?" (1990).

Contents

1. INTRODUCTION

When (re)designing or evaluating complex man-machine systems, human factors specialists generally recognize the importance of assessing the cognitive load the system imposes on the operator. One of the procedures they frequently use for an objective and quantitative assessment of cognitive load is the reserve or spare capacity task methodology. The use of this methodology is based on the assumption that the operator has a limited capacity to process task-relevant information. Increases in task difficulty are assumed to lead to increases in processing resource expenditure up to a point where the capacity limit is exceeded and degradations in performance result. The objective of the methodology is to measure the spare capacity available to the operator while carrying out the task of primary interest. To this end, the operator is required to perform an additional or a secondary task, along with the primary task. Furthermore, in most applications, the operator is instructed to maintain primary-task performance at the expense of the secondary task. Then, a degradation in secondary-task performance can be seen as an index of the residual capacity associated with the primary task. For an extensive discussion of the methodology and its underlying assumptions, the reader is referred to reviews of O'Donnell and Eggemeier (1986), Ogden, Levine, and Eisner (1979), or Williges and Wierwille (1979).

In the literature, a multitude of secondary tasks has been employed within the spare capacity paradigm. The most commonly used classes of tasks are: choice reaction time, memory, monitoring, and tracking. Given this diversity, care must be taken in selecting an appropriate task for the particular application at hand. Fortunately, several authors have listed a number of criteria which may be used to guide this selection (e.g., Knowles, 1963; O'Donnell and Eggemeier, 1986). One main criterion is sensitivity. This criterion says that the secondary task should have the capability to discriminate significant variations in the cognitive load imposed by the primary task. Empirically, sensitivity can be tested by manipulating primary-task difficulty and observing the effect of it on secondary-task performance. In order to maximize sensitivity, a variety of guidelines has been formulated. These guidelines include using a secondary task with one (or more) of the following properties: (1) continuity (i.e., imposing an ongoing demand on the information-processing system); (2) stability (i.e., not showing strong variations in the demands because of significant practice or learning effects); and (3) representativeness (i.e., having the maximum possible overlap with the demands of the primary task).

Another major criterion against which the appropriateness of a secondary task can be evaluated is interference with the primary task. Ideally, the secondary task should in no way change primary-task performance. Performance on the primary task should be completely secured when the secondary task is introduced. This is because under dual-task conditions it is assumed that all the consequences of exceeding the capacity limit show up only in secondary-task performance. When interference occurs, secondary-task performance decrements do not represent a pure index of the reserve capacity associated with the primary task. Then, clear interpretation of the results becomes difficult. Interference with the primary task also puts severe restrictions on the practical applicability of the methodology. In some settings, especially in operational environments, significant degradations in primary-task performance are not acceptable because of safety risks. Empirically, interference can be tested by comparing performance on the primary task when carried out concurrently with the secondary task with performance on the primary task when carried out alone (i.e., under control conditions). In an effort to minimize interference, several recommendations have again been made. These recommendations include using a secondary task which is self-paced (i.e., presented at a rate determined by the subject him/herself) or which is embedded (i.e., already exists as a component of the subject's normal repertoire of activities). Additional criteria have been formulated for guiding the selection of an appropriate secondary task, such as ease of implementing and operator acceptance. For a more complete and detailed list of criteria, the reader should consult the review of O'Donnell and Eggemeier (1986).

It should be noted that the (relative) importance of the criteria for secondary-task selection is highly dependent on the purpose of applying the spare capacity task methodology. A secondary task which satisfies the purpose in one application need not be satisfactory for another application. For example, in a laboratory environment a high degree of interference with the primary task may be more acceptable than in a field situation. Furthermore, although in a given application some of the currently available tasks may come close to meeting many of the relevant criteria, only a few, if any, may meet them all. Because of reasons like these, the search for secondary tasks which fulfil the usually highly specific requirements of the particular application at hand still continues.

Although not explicitly referring to the spare capacity methodology, Bromme and Wehner (1987) have suggested to use think-aloud measures as an index of the cognitive load afforded by the primary task. Simply speaking, thinking aloud requires the subject to overtly verbalize the thoughts he/she is engaged in when, at the same time, carrying out the primary task. In the usual applications, the verbalizations the subject produces are used to infer the cognitive processes that go on during task performance. With this goal in mind, the think-aloud technique has been extensively and successfully applied in a diversity of task domains of which physics, mathematics, process control, and computer programming are only a few examples. The point now is whether thinking aloud is also an appropriate technique for measuring the cognitive load associated with a primary task in terms of secondary-task performance.

The theoretical basis for using thinking aloud as a secondary task is provided by the model of verbal reporting developed by Ericsson and Simon (1984). Within the framework of this model, thinking aloud involves two kinds of information-processing activities: those directed at the primary task and those related to the act of verbalization. It is assumed that, at least under some conditions, both activities impose a certain demand on short-term memory (STM). More specifically, thinking aloud is thought of as verbalizing (a portion of) the information that is entered into STM by the cognitive processes used in the primary task. Here, STM is seen as a unitary mechanism which has a limited capacity for temporarily storing the information to be processed. From this it follows that the larger the amount of STM capacity required by the primary task, the less remains for verbalizing concurrently. One may thus predict that with an increase in primary-task load, think-aloud performance will deteriorate.

When used as a secondary task, thinking aloud differs in one fundamental way from the tasks which are normally used for this purpose. While the usual secondary tasks may be seen to be entirely separate and distinct from the primary task, thinking aloud is directly related to it. In fact, it is subordinate to and closely follows the cognitive processes used in the main task. In Ericsson and Simon's view, this is because thinking aloud involves only the concurrent verbalization of information already produced by the primary cognitive processes. On the one hand, this tight coupling suggests that thinking aloud will be highly sensitive to small but significant variations in the cognitive load of the primary task. In particular the following two properties seem to contribute to this: continuity and stability. As noted previously, properties like these are especially recommended to increase sensitivity. On the other hand, the coupling with the main task also seems to increase the likelihood of interference. As a matter of fact, Ericsson and Simon in their model of verbal reporting explicitly point out under what conditions this is likely to occur. For this purpose, they distinguish among three levels of verbalization. At the first level, the information to be reported about is already available in STM and verbally encoded. This information can therefore directly be verbalized without imposing additional demands on STM. In this case, thinking aloud will in no way interfere with primary-task performance. That is to say, neither the course and structure nor the speed of the primary cognitive processes will undergo significant changes. At the second level of verbalization, the requested information is also available in STM but in a non-verbal (e.g., visual) code. This information must therefore be translated into a verbal code when verbalizing it. Since such a translation makes at least modest demands on STM, thinking aloud has the effect of slowing down the primary cognitive processes. Nevertheless, the course and structure of these processes will not be affected. At the third level of verbalization, the required information is not directly available in STM. Instead, this information must first be generated by means of additional interpretative processing, like filtering or inference, before it can be verbalized. In this case, thinking aloud may change the primary cognitive processes fundamentally. Thus, Ericsson and Simon's view may be summarized as follows. The larger the amount of intermediate information processing intervening between the primary task and thinking aloud, the more severe the interference.

After having dwelled upon the theoretical notions surrounding the use of thinking aloud as a secondary task for measuring cognitive load, we now turn to the available empirical evidence. It should be noted, however, that in the literature hardly any study is reported which has specifically been designed to investigate this issue empirically and in a systematic way. An exception is the study carried out by Bromme and Wehner (1987). As a consequence, there is no extensive database dealing with thinking aloud as a secondary task. Nevertheless, a reasonably number of studies has been published which provide some data bearing on the sensitivity of thinking aloud. Since sensitivity is an important criterion to be fulfilled for secondary-task measurement of cognitive load, we will consider this research more closely.

In the research that has been done, it is examined whether think-aloud performance fluctuates as a function of primary-task difficulty. With this goal in mind, a number of subjects is required to verbalize their thoughts while carrying out a task the difficulty of which is experimentally manipulated or determined *post hoc* on the basis of the subjects' performance. From the verbalizations the subjects produce one or more measures are derived to capture the temporal or qualitative aspects of thinking aloud. These measures are then related to the different levels of task difficulty. In this research, use has been made of all kinds of tasks, such as anagrams (Deffner, 1984), the N-term series problem (Kempkensteffen, 1987; Ohlsson, 1980), geometric puzzles (Deffner, 1984), problems in physics (Simon and Simon, 1978), and determining the course order for a dinner (Bromme and Wehner, 1987). Furthermore, various measures of think-aloud performance have been used. Nevertheless, it is common practice to include at least the rate of verbalization, defined here as the number of speech utterances (e.g., words or syllables) produced per unit of time (e.g., minutes).

Unfortunately, the results of the studies under consideration show a disappointing lack of consistency in results. In some cases, increasing task difficulty decreased think-aloud performance (Bromme and Wehner, 1978), but in many other cases (hardly) no effect has been observed (Deffner, 1984; Deffner and Ericsson, 1985; Ohlsson, 1980; Simon and Simon, 1978). Given the latter results, one might conclude that thinking aloud is not a particularly reliable secondary task for measuring variations in cognitive load. And yet, the present data are not fully adequate to draw such a conclusion. Apart from the fact that the available evidence is rather modest, the studies conducted so far suffer from a number of methodological imperfections which pose serious problems in interpreting the results obtained.

First, many studies fail to determine whether thinking aloud interferes with performing the primary task. As noted previously, when interference occurs, thinking aloud does not represent a pure index of the reserve capacity associated with the primary task. With respect to this point, it is noteworthy that in the literature several studies are reported where thinking aloud interferes substantially with the main task. In these cases, the primary cognitive processes seem to change fundamentally because of the requirement to verbalize them concurrently (e.g., Russo, Johnson, and Stephens, 1989). Furthermore, it appears that with thinking aloud a large variety of tasks take considerably more time to perform. Among these tasks are not only those which are supposed to be executed using visual codes, but also those supposedly involving the use of verbal codes (e.g., Rhenius and Deffner, 1990). So, concurrent verbalization seems to exert a general slowing down effect on the primary cognitive processes.

Apart from failing to control for interference, practically all the studies fail to take into account that the primary task might be carried out in qualitatively distinct ways, i.e., by following basically distinct strategies. Especially when the task is relatively complex, there may be a variety of strategies available. In that case, one should allow for the possibility that one strategy is changed for another when task difficulty increases. And indeed, the literature provides numerous examples where subjects in response to an increase in task difficulty adopt a less demanding strategy (e.g., Sperandio, 1978). In the situations considered here, such a strategy shift could have the effect that performance on the primary task deteriorates while think-aloud performance does not. Furthermore, when there are many alternative strategies for doing a task, one should also allow for the possibility that the dimension of difficulty being selected is not effective for each possible strategy. That is to say, it may very well be that during the task some strategy is adopted the demands of which do not increase with higher levels of the chosen dimension of difficulty. In the situations under consideration, increases along a (partly) ineffective dimension of primary-task difficulty could also have the effect that think-aloud performance does not deteriorate accordingly.

Given the methodological problems associated with many of the studies conducted so far, it seems premature to dismiss thinking aloud as an appropriate secondary task. Therefore, the goal of the present study is to provide more direct evidence with respect to this issue. In particular, the study examines whether think-aloud performance, as indexed by the rate of verbalization, is sensitive to variations in the difficulty of the primary task. In trying to answer this question, the present methodology improves on past research in two important ways. First, the degree of interference is evaluated. This is done by comparing performance on the primary task when combined with thinking aloud with performance on the primary task when carried out alone, i.e., in silence. Secondly, strategy use in the primary task is explicitly controlled for. This is realized by contrasting the think-aloud performance of the following three groups of subjects:

1. those who despite increases in task difficulty stick to their strategy, thereby facing higher cognitive demands;
2. those who in response to increases in task difficulty adopt a cognitively less demanding strategy;

3. those adopting a strategy for which the chosen dimension of task difficulty is not effective.

For the sake of simplicity, these strategy groups will henceforth respectively be referred to as: the persisting subjects, the yielders and the insensitive subjects. The expectation is that the persisting subjects will show a larger decrease in think-aloud performance than the yielders. It is also expected that the insensitive subjects will show no change in think-aloud performance.

2. METHOD[2]

In the experiment, two types of data were collected. One type of data referred to the subjects' primary-task performance and to their think-aloud performance (2.1), the other type of data referred to the subjects' strategies in the primary task (2.2).

2.1. Collecting performance data

Subjects and design

Twenty-four male students from the Eindhoven University of Technology served as subjects. They were paid per hour for their participation.

Each subject performed the primary task in two conditions: (1) while thinking aloud; and (2) in silence. In both conditions, the difficulty of the task varied according to two levels, referred to as easy and difficult. To control for possible order effects of condition, ABBA counterbalancing was employed. This means that each subject underwent the two conditions first in one order and then in the reverse order. This design, however, does not ensure that order effects are balanced out if asymmetric transfer occurs. To cope with this, different condition orders were incorporated in the design and each subject was allocated to one of these. The following two orders were included: 1221 and 2112. Here, the consecutive numbers in a sequence refer to the successively administered conditions with each administration involving two experimental problems to be solved. For example, a subject assigned to the second condition order subsequently solved two problems in silence, four problems requiring thinking aloud and two problems in silence. Fortunately, condition order did not have any systematic effect on task performance. This factor could therefore be excluded from the analyses of primary interest. The four problems administered first alternately consisted of an easy and a difficult problem and for the second four problems this order in level of difficulty was reversed.

Before being administered, the complete set of 12 experimental problems was first randomly split up into three subsets with each subset including two problems of the same level of task difficulty. Then, the subsets were assigned to the experimental conditions

[2] Portions of this chapter have been copied from Brinkman (1990).

according to three different ways. For each way of problem assignment, one subset was meant for thinking aloud and another for silence. The three ways of problem assignment were designed so that the subsets were counterbalanced over the conditions. Finally, each subject was randomly allocated to one way of problem assignment with the constraint that each way involved eight subjects.

Primary task

The primary task employed has been patterned after the fault diagnosis task developed by Brooke, Duncan, and Marshall (1978). Unfortunately, the task of Brooke *et al.* did not seem to have the potential to induce the diversity of strategies required for the present experiment. Therefore, several changes were introduced in that task, although its main component, the fault-indicator matrix, was retained. The modified task involves identifying a fault in a hypothetical system in which a number of interrelated variables is operating. In the system, one of N potential faults occurs and affects the available variables. There are two types of variables, namely hidden and indicator variables. The precise difference between these two variable types is explained later. In the following, Fi represents fault i, Xj represents hidden variable j and Yk indicator variable k; the subscripts i, j and k stand for a code number which can take any value between 1 and N, inclusive.

The interrelations between the variables are laid down in a set of logical expressions an example of which is given in Table 2.1. Each expression has in front of its equation sign two input variables separated by a logical operator and behind its equation sign an output variable. Note that the very same variable can be found in two or more expressions and can act as an input variable in one expression and as an output variable in another.

Table 2.1. An example of a set of logical expressions describing the system to be diagnosed.

[1]	X6	AND	Y1	=	Y2	[4]	X3	EQV	Y1	=	Y4
[2]	Y1	OR	Y2	=	X5	[5]	Y6	AND	Y3	=	X2
[3]	Y5	OR	X1	=	Y6	[6]	Y2	AND	X4	=	Y5

Each variable takes one of two possible values, namely 1 or 0. In a given expression, the value of the output variable is determined by:

1. the values carried by the input variables;
2. the logical operator.

Three logical operators are used: AND (logical conjunction, the output variable gets the value of 1 if both input variables are 1, otherwise it becomes 0), OR (logical disjunction, the output variable takes the value of 1 if at least one input variable is 1, otherwise it will be 0), and EQV (logical equivalence, the output variable gets the value of 1 if both input variables have the same value, whether 1 or 0, otherwise it becomes 0).

The working of the system to be diagnosed is as follows (see Figure 2.1). If a fault with a particular code number occurs, the value of the hidden variable with the same code number changes from 1 to 0, whereas the values of all the other hidden variables remain 1. Thus, each possible fault uniquely determines the values of the hidden variables. These in turn determine the values of the indicator variables. The way in which this occurs is dictated by the set of logical expressions being operative. That is to say, the values of the

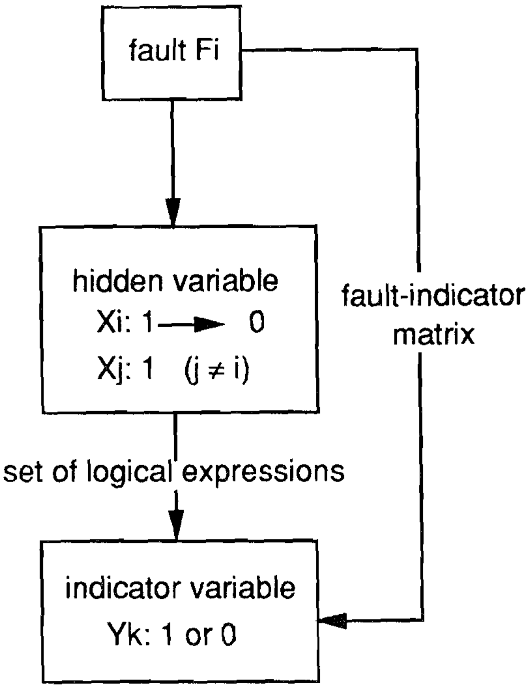


Figure 2.1. Flow diagram of the fault diagnosis task employed. See text for explanation.

hidden variables generated by a particular fault have to be entered into the logical expressions in order to compute the values of the indicator variables belonging to the fault. Of course, such computations can be carried out for each possible fault. The results thereof may then be arranged in the form of a fault-indicator matrix. A row in this matrix indicates what values the different indicator variables take if a particular fault occurs. A column indicates the values of one particular indicator variable with the occurrence of the different faults. An example of a fault-indicator matrix is portrayed in Table 2.2. This matrix has been derived from the set of logical expressions given in Table 2.1.

Table 2.2. The fault-indicator matrix associated with the set of logical expressions presented in Table 2.1.

		Indicator variable					
		Y1	Y2	Y3	Y4	Y5	Y6
Fault	F1	1	1	1	1	1	1
	F2	1	1	0	1	1	1
	F3	1	1	1	0	1	1
	F4	1	1	1	1	0	1
	F5	0	0	1	0	0	1
	F6	1	0	1	1	0	1

At the start of a problem, the subject is given a set of logical expressions describing the system to be diagnosed. He also receives a partly completed fault-indicator matrix associated with the set. In order to gain information on the actual fault present in the system, the subject can test an indicator variable or a potential fault. The subject tests an indicator variable by giving a terminal command of the following form:

```
Yk <return>
```

Upon entrance of such a command, the subject is provided with the actual value of the indicator variable tested. The subject can test a potential fault by entering a terminal command of the form:

```
Fi <return>
```

Having entered such a command, the subject is informed whether the potential fault he tested is the actual one or not. If it is not, the subject can continue making tests and if it is, the problem is solved by him.

Execution of the tests occurs in consecutive trials. For a test on an indicator variable one trial is charged and for a test on a potential fault three trials. If the subject tests a non-existent indicator variable or fault, or enters a syntactically improper command, he will be informed so. Then, the trial number is not raised. Figure 2.2 illustrates the manner in which the tests are successively performed.

```
test        Y5
test        Y1
diagnosis   F4
test        Y4
diagnosis   F6
```

Figure 2.2. Format of the text display located at the subject's left hand. The display presents an historical overview of the successive tests the subject made during the fault diagnosis task. The subject diagnosed the system described by the set of logical expressions given in Table 2.1.

The fault diagnosis task described here is realized by applying a Minimum Information Feedback (MIF) procedure (after Johnson, 1978). This procedure has the effect of maximizing the number of tests required to find the fault by minimizing the information value of the feedback provided. When using the MIF procedure, the computer program controlling the task continuously keeps track of all the faults which, given the information provided so far, are still logically tenable. No matter which test the subject asks for, the computer program presents that kind of feedback which keeps this set of possible faults as large as possible. In fact, the fault designated as the actual one is not predetermined at the start of the problem but is the last one which the subject eliminates logically. This procedure guarantees that no problem will be abruptly terminated because of a lucky guess of the subject. Indeed, application of the MIF procedure in the present experiment insures that the subject needs at least four tests in order to identify the fault. This maximization of the number of tests required contributes to a valid identification of the subject's strategy.

Strategies

By combining the results of normative modelling attempts presented in the literature (Brooke, Duncan, and Cooper, 1980) with an analysis of verbal reports obtained in a pilot experiment, we identified for the fault diagnosis task described above a number of possible strategies. In the following, these task strategies are described from a more logical and a more psychological point of view.

Logical strategy description. There appear to be two basically different strategies according to which the fault diagnosis task can be performed, namely hunting and scanning.

When involved in hunting, the subject first selects a more or less random indicator variable and tests its value. Next, he enters the value obtained in the logical expressions being operative and tries to derive what hidden variable takes the value of 0. If the subject fails in finding this variable, he then chooses another indicator variable for a test and repeats the procedure. If, however, he does succeed in it, he then performs a test on the fault which affected the hidden variable found and accordingly solves the problem. It should be pointed out that, prior to testing an indicator variable, the subject fails to examine whether the outcome might be used to eliminate one or more faults which are still logically tenable. It is only after the subject has performed the test that he checks whether the actual value obtained allows such an elimination. Note that a feasible fault can be eliminated as soon as the hidden variable with the same code number appears to be 1. In short, when involved in the hunting strategy, the subject continues in making tests on more or less randomly selected indicator variables until the logical computations he performs on the results thereof reveal what hidden variable is equal to 0.

With the scanning strategy, the subject explicitly looks for an indicator variable which has the potential to reduce the set of faults which are logically feasible. The characteristic feature of an indicator variable allowing such a reduction is that it does not take one and the same value with all the faults making up the feasible fault set. The subject determines whether a given indicator variable possesses this feature by deriving from the provided set of expressions what possible values this variable takes under the feasible faults. When making these derivations, the subject actually completes the presented fault-indicator matrix. In particular, he fills in the empty cells of the column which is associated with the indicator variable under consideration. Having found the desired variable, the subject performs a test on it and, on the basis of its actual value, he removes those faults from the feasible set which are no longer tenable. Once the subject has reduced the feasible set so as to contain only one fault, he tests that fault and accordingly solves the problem. In sum, the scanning strategy involves completing the fault-indicator matrix in order to identify those indicator variables where the faults which are logically feasible produce different results.

When comparing the two basic strategies, it appears that they differ in the extent to which at any particular moment use is made of the available information. As will be evident from the preceding, the amount of information utilized is smaller when adopting the hunting strategy than the scanning strategy. It thus seems appropriate to qualify hunting as less thorough than scanning.

It should be kept in mind that a subject can perform the fault diagnosis task by using a strategy which differs in one or more respects from the basic strategies described here. When this occurs, the subject is said to follow an indefinite strategy. An example of an indefinite strategy is one in which all the available indicator variables are indiscriminately tested before the results obtained are used for deriving the actual fault. It should also be taken into account that the subject, while doing the task, can shift from one strategy to another. For instance, the subject may first test randomly selected indicator variables and adjust the set of feasible faults upon each test performed. Having reduced the set so as to contain a small number of faults, the subject may then look for indicator variables which he, prior to testing them, checks for their capacity to eliminate the remaining faults. So in this example, the subject starts by employing the hunting strategy and then turns to the scanning strategy.

Psychological strategy description. In the hunting as well as the scanning strategy a fault elimination procedure is adopted which involves performing mental operations on the presented set of logical expressions and retaining the intermediate results of these operations for some time. Therefore, the two strategies seem to call for the same kinds of cognitive processes, viz. logical reasoning and memory processes like rehearsal.

Nevertheless, they differ considerably in the quantity of resources required for these processes. This is because of their difference with respect to the extensiveness of the fault elimination procedure applied. In the hunting strategy, each indicator variable being tested is simply selected at random. In addition, each test value obtained is immediately entered into the logical expressions which have to be passed through only once in order to eliminate the impossible faults. This contrasts with the scanning strategy in which, prior to each test performed, a search is made for an indicator variable which takes different values under the feasible faults. During the search, several passes have to be made through the logical expressions, namely once for each indicator value to be derived. Furthermore, the derived values must be retained until the appropriate indicator variable has been found. It thus appears that the elimination procedure adopted in the scanning strategy is more elaborate than the one applied in the hunting strategy. The implication is that scanning, in comparison with hunting, depends more on logical reasoning and places higher loads on memory. Thus, the scanning strategy is more cognitively demanding than the hunting strategy.

Differences in the processing requirements of the distinct strategies will manifest themselves as differences in task performance. Specifically, when employing the hunting strategy, a relatively rapid sequence of tests is performed and no attention is paid to whether each test is redundant or not. So with this strategy, the tests are made quickly after another, but many are needed to find the fault. In the scanning strategy, on the other hand, a test is selected because of its informative value and is only made after deliberate and slow reasoning. Hence, finding the fault according to this strategy requires a small number of tests, but each test takes a considerable amount of time.

Instructions

Two sets of written instructions were prepared, in particular a task and a think-aloud instruction.

Task instruction. This instruction explained the nature of the primary task and the use of the computer terminal. It also described the two basic task strategies although the subject was left free to follow the strategy he preferred. The subject was further instructed to complete the task in a minimum of test trials and at his own pace. So, accuracy rather than speed was stressed. In order to prevent the subject from making an overwhelmingly large number of tests and using an excessive amount of time, he was told to solve the task in less than 50 trials and within 15 minutes, otherwise the task would be terminated. A pilot study indicated that these constraints provided ample opportunity for completing the task appropriately. Two paper and pencil exercises were added to the task instruction for the purpose of testing the subject's understanding of the task.

Think-aloud instruction. In this instruction, which had been adapted from Ericsson and Simon (1984), thinking aloud was explained to the subject as verbalizing everything he was thinking of from the moment the task began till its end. The instruction encouraged him to talk aloud constantly but dissuaded him from planning beforehand what to say and from explaining what he said.

Stimulus materials

Fifteen problems to be administered by a computer were developed. Three problems were utilized for practice. These were especially meant to train the subject in verbalizing his thoughts and in on-line task performance. The other 12 problems served experimental purposes.

For each problem, the computer generated a different set of logical expressions. When producing a set of expressions, the following constraints were regarded. A complete set invariably consisted of six logical expressions which were made up of six different hidden variables, six different indicator variables, and three types of logical operators. Each expression had the same construction, consisting of one hidden variable, two indicator variables, and one logical operator. The hidden variable to be allocated to a given expression was always selected at random. However, this selection was done so that each expression in the set contained another hidden variable. Hence, each hidden variable occurred only once in the complete set. The selection of the first indicator variable to be assigned to a particular expression proceeded in a similar way. Furthermore, all the indicator variables were equally likely to serve as the second indicator variable in an expression. This with the constraint that the same indicator variable could not be assigned twice to a particular expression. As a result, the number of times each indicator variable occurred in the complete set of expressions could vary from one up to and including six. There were three distinct locations in an expression, two before the equation sign and one behind it. What location a particular variable occupied was randomly determined. The logical operator being selected for an expression was equally likely to be AND, OR, or EQV.

Apart from the above-mentioned restrictions, each set of logical expressions was also constrained in the following three ways. First, when solving a set of expressions for any potential fault, each indicator variable would yield one particular value. So, each cell in the fault-indicator matrix contained only one value. Secondly, the set of indicator values produced by each fault was unique. Hence, any row of values in the fault-indicator matrix differed from any other row. Finally, each indicator variable was associated with another typical set of values under the various faults. So, any column of values in the matrix was different from any other column.

Task difficulty was manipulated by varying the amount of information available in the fault-indicator matrix. That is to say, the problems to be administered differed with respect to the number of empty cells in the fault-indicator matrix presented. Here, a problem was conceived of as being more difficult if its fault-indicator matrix contained a larger number of empty cells. This was accomplished by deriving the fault-indicator matrices from the generated sets of logical expressions and by leaving 12 or 24 cells in each matrix blank. Thus, two different levels of task difficulty were created. What cells in a matrix were not filled in was determined at random, though subject to the constraint that each row and each column contained the same number of empty cells. Of the practice problems, there was one of which the fault-indicator matrix consisted of 12 empty cells. The fault-indicator matrices of the other two practice problems had 24 empty cells. Of the experimental problems, half had a matrix consisting of 12 empty cells and the other half had a matrix with 24 of such cells. The fault-indicator matrices thus prepared were drawn on separate sheets of paper.

It should be realized that the effectiveness of the chosen dimension of task difficulty depends upon the type of strategy employed. Remember that the scanning strategy is explicitly directed at the use of the fault-indicator matrix. This does not hold for the hunting strategy. Consequently, experimental manipulation of the number of empty cells in a fault-indicator matrix should indeed affect performance when adopting the scanning strategy but should exert no systematic performance effect with the hunting strategy.

Technique and apparatus

A subject was tested in isolation in a sound-attenuated room. In this room, he sat behind a table upon which two terminals were placed. One terminal was located in front of the subject and the other at his left hand. The first terminal consisted of a text display with an ASCII keyboard connected to it. The second terminal was a text display without a keyboard. Both terminals were driven by a computer which was programmed to control the fault diagnosis to be executed by the subject. The terminal in front of the subject served two purposes. First, it displayed the set of logical expressions being operative during the task. Second, it could be used by the subject for obtaining test information from the computer. The terminal at the subject's left hand was used to present an historical overview of all the tests the subject successively performed. On the table, at the subject's right hand, was a small reading desk upon which the set of papers with the fault-indicator matrices was lying. The computer recorded the sequence of tests the subject made and the response times he displayed.

During the experiment, the subject was monitored by the experimenter who sat in a room next to the test room. From his room, the experimenter controlled the presentation of the task by interacting with the computer through another terminal. Subject and experimenter communicated with each other through an intercom system. A complete record was made of the subject's verbalizations. For this purpose, a microphone was hung around the subject's neck and connected to a tape-recorder in the experimenter's room. The recorder was operated by the experimenter.

Procedure

A subject was tested individually and participated in one experimental session. A session consisted of a practice and an experimental phase.

Practice phase. In this phase, the subject first studied the task and verbalization instruction. He then made the two paper and pencil exercises added to the task instruction. Having completed the exercises, he was given feedback about his performance. Next, the subject solved the three computer-administered practice problems. The first problem had to be solved in silence, the second one required concurrent verbalization and the third one retrospective verbalization. Upon completion of these problems, the subject received feedback about the quality of his verbalizations. In the practice phase, the subject was free to ask any relevant question, but he was not given any information about the goal of the experiment.

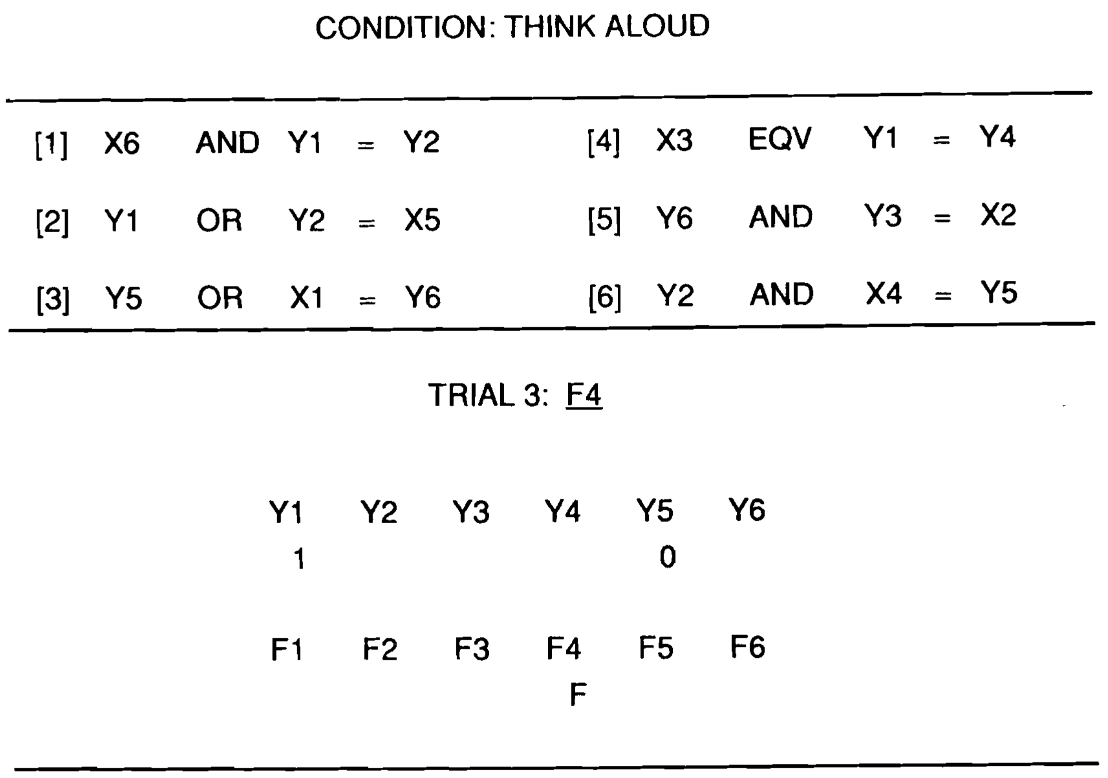
Experimental phase. In this phase, the subject solved the eight experimental problems presented by the computer. Each problem had to be carried out in one of the two conditions. The sequence in which the conditions were administered proceeded according to one of the two possible orders mentioned earlier.

At the start of each problem, the text display located in front of the subject indicated the experimental condition. This indication remained on the display for the time the subject solved the problem presented. The experimenter saw to it that the subject obeyed to this. That is to say, when the subject was asked to think aloud and remained silent for more than about 10 seconds, the experimenter would persuade him to resume verbalization. Conversely, when silence was requested from the subject and he in spite of that started verbalizing during the problem, the experimenter would warn him to be silent.

During each problem, the text display in front of the subject was divided into an upper and a lower section (see Figure 2.3). The upper section contained the set of logical expressions being operative. The lower section was made up of two rows of figures, one row displaying the available indicator variables and the other row showing the potential faults. As soon as in the latter section the current trial number appeared, the subject could test an indicator variable or a fault by activating the keyboard. Having entered the desired test, the outcome was shown on the appropriate position below the relevant row. Test outcomes obtained in previous trials were not erased from this display. In addition, the text display at the subject's left hand kept a complete record of the successive tests the subject performed (see Figure 2.2).

When performing a problem, the subject was allowed to consult the sheet of paper containing the fault-indicator matrix which corresponded with the set of logical expressions presented. Furthermore, the subject could refer to a paper explaining the meaning of the logical operators being applicable. However, he was not permitted to use paper and pencil.

A problem ended when the subject succeeded in solving it. A problem was also terminated if the subject exceeded the maximum number of 25 test trials permitted or the time limit of 20 minutes. However, this was not done before the moment the subject performed the test which went beyond (one of) the limits.



CONDITION: THINK ALOUD

[1]	X6	AND	Y1 =	Y2		[4]	X3	EQV	Y1	=	Y4
[2]	Y1	OR	Y2 =	X5		[5]	Y6	AND	Y3	=	X2
[3]	Y5	OR	X1 =	Y6		[6]	Y2	AND	X4	=	Y5

TRIAL 3: F4

Y1	Y2	Y3	Y4	Y5	Y6
1				0	

F1	F2	F3	F4	F5	F6
			F		

Figure 2.3. Format of the text display located in front of the subject. The command entered by the subject in the current trial has been underlined. See text for explanation.

Performance measures

To describe a subject's performance on the primary task, the following two measures were determined for each problem the subject solved: the time taken and the number of test trials needed. To verify the validity of the procedure followed for identifying the subjects' strategies, the following two measures were determined as well: the time taken per test and the number of redundant tests (these are tests which do not result in a reduction of the set of feasible faults).

The subject's think-aloud performance was described by the rate of verbalization. To obtain this measure, the tape-recorded verbalizations, produced by the subject in the think-aloud condition, were transcribed into typewritten form. In making these transcriptions, essentially all the audible speech utterances were written down. No corrections of grammar were made and punctuation was omitted. Once written down, the transcriptions were carefully edited for accuracy. In this way, all the protocols from a total number of 23 subjects were transcribed. A number of the protocols from the only remaining subject could not be transcribed because of poor tape-recording quality. This subject was excluded from the analyses described below. The rate of verbalization was determined by calculating for each transcribed verbal protocol the number of syllables produced per minute.

2.2. Collecting strategy data

In his study, Brinkman (1990) reports about an algorithm specifically developed to identify the strategies the subjects follow in the task employed. In this algorithm, the strategy classifications are made on the basis of non-verbal behavioral recordings, in particular the tests the subjects successively enter to solve the task. It appeared, however, that these recordings were not fully adequate to be used for strategy identification as the algorithm frequently failed to distinguish one strategy from another. In the present study, it was therefore decided to have a number of coders identify the subjects' strategies from their verbal protocols.

Coders

Six subjects, who had previously taken part in the experiment, served as coders. They were paid for their participation on an hourly basis. The coders were randomly divided into two groups of equal size. Hereafter, the groups will be referred to as group 1 and 2.

Coding materials

Since the analysis of a verbal protocol may be rather laborious, it was decided to have each coder analyze only a portion of all the verbal protocols being transcribed. For this purpose, two sets of protocol transcriptions were selected. Each set consisted of the protocols from a completely different group of subjects. All the protocols which these subjects had generated in the think-aloud condition (i.e., four per subject) were included. Thus, a set was made up of a total number of six (subjects) times four (protocols per subject) is 24 protocols. Each coder of group 1 received one set and each coder of group 2 the other. Care was taken that a coder did not get any of the verbal protocols which had been produced by himself during the experiment.

The order in which the selected verbal protocols were presented was randomized for each coder. This was done to reduce the possibility that the coder, while analyzing a subject's verbalizations contained in the current protocol, would infer from a preceding protocol what the subject might have been thinking.

Coding instructions

A written instruction was prepared containing detailed information being essential for the task of analyzing verbal protocols. In the first part of this instruction, a complete description was given of the different fault diagnosis strategies together with typical examples. This part essentially consisted of the logical strategy specification presented above. No reference was made to the psychological implications of the strategies. The second part described at great length the standardized way in which the verbal protocols were to be encoded into these strategies. This part in particular consisted of five examples of carefully encoded protocols.

In analyzing a protocol, the coder was instructed to proceed according to the following two steps. First, he had to derive which diagnostic tests the subject successively made. Then, he had to encode each test, with the exception of the one performed on the faulty component, into one of the distinct strategies. The coding categories to be used were: hunting, scanning, and indefinite. The coder was emphatically stimulated to base his strategy classifications on the subject's overt verbalizations rather than on plausible inferences as to what the subject might have been thinking.

Another series of five verbal protocols was compiled for the purpose of providing the coder with concrete practice in protocol analysis. In an additional set of papers, these protocols were encoded in a way as outlined in the instruction. This set would serve feedback purposes.

Coding procedure

A coder took part in at least four sessions and at most seven. He completed the sessions over a period ranging between five and ten days. In each session, the coder worked individually and at his own pace. In the first two sessions, the coder received extensive training in protocol analysis. In order to refresh his memories, he first read again the instruction of the fault diagnosis task and then once more solved a number of computer-administered diagnostic problems. Thereupon, he studied the instructions describing the task of protocol analysis. He indicated when he had completed this instruction and then encoded the verbal protocols that made up the practice set. After having finished this work, he was given the set of papers describing the correct protocol encodings with which he was to compare the encodings he had made. Whenever he observed differences in it, he would discuss these with the experimenter to find the source. From the third session on, the coder worked through the set of verbal protocols selected to be analyzed.

Strategy identification

Differences between the coders in the strategy encoding given to a particular test were resolved by applying a majority principle. That is to say, the encoding selected for further analysis was always the one given by at least two of the three coders making up a group. If there was no majority for one of the strategies, the encoding indefinite would be used.

After the individual tests of a problem had thus been coded, the problem as a whole was classified as hunting, scanning or indefinite. This was done according to which strategy category contained the majority of the encoded tests. If there was no majority, the classification "mixed" was used. Following this, each subject was classified. Specifically, if the two easy problems a subject had solved received the same strategy classification, the subject was identified accordingly, i.e., as a hunter, as a scanner, as indefinite or as mixed. If these problems differed with respect to the given classification, the subject was identified as mixed. For the difficult problems, the same procedure was followed. Thus, each subject was classified twice, once for the low level of task difficulty and once for the high level.

3. RESULTS

The first step in the analyses was to verify whether the procedure followed to identify the subjects' strategies was adequate (3.1). After having found evidence for this being the case, the tests of primary interest were carried out (3.2).

3.1. Adequacy of strategy identification procedure

To be adequate, the strategy identification procedure should produce data that are both reliable and valid.

Reliability of strategy data

The reliability of the strategy data was evaluated by determining inter-coder agreement. This was done by calculating for each pair of coders from the two groups involved the correspondence between the strategy encodings extracted from the verbal protocols. For this purpose, all the encodings belonging to one pair of coders from a given group were cast in the form of a contingency matrix an example of which is shown in Table 3.1. An inspection of the individual matrices revealed no systematic differences between the coders in their agreement about the different strategies.

To measure the degree of agreement between the coders, Cohen's kappa (Cohen, 1960) was applied to each contingency matrix obtained. This measure gives the proportion of agreement after chance agreement has been removed from consideration. Across all pairs of coders, kappa ranged from 0.79 to 0.92 with a mean of 0.86 (N varied from 123 to 126 with a mean of 124.17). The mean kappa indicates that the inter-coder agreement was on the average almost perfect.

As described above, the two groups of coders analyzed only a portion of all the transcribed verbal protocols. Given the high inter-coder agreement, it seemed reasonable to have only one person analyze all the remaining protocols. The person carrying out these additional analyses was the author himself. He thereby followed the same procedure as the coders had.

Table 3.1. Contingency matrix based on the concurrent verbalization encodings derived by coders A and B. Parenthetical values are proportions.

		Coder B			
	Strategy	HU	SC	IN	TOTAL
Coder A	HU	31 (0.25)	5 (0.04)	0 (0.00)	36 (0.29)
	SC	1 (0.01)	81 (0.66)	0 (0.00)	82 (0.67)
	IN	2 (0.02)	0 (0.00)	3 (0.02)	5 (0.04)
	TOTAL	34 (0.028)	86 (0.70)	3 (0.02)	123 (1.00)

Key. HU: hunting; SC: scanning; IN: indefinite.

Validity of strategy data

The validity of the strategy data was evaluated by examining whether the subjects displayed the behavioral characteristics implied by the task's psychological strategy description presented above. According to this description, the different strategies should be accompanied with distinctive behavioral features. Specifically, the hunting strategy would be associated with a large number of tests possibly being redundant and each attempted in a short time. The scanning strategy, on the other hand, would be associated with a few number of informative tests each requiring a lot of time. Suppose that these relationships, specified beforehand, were really reflected in the data obtained afterwards. This then would indicate that the data resulting from the strategy identification procedure were valid. Therefore, the hunters and the scanners were compared with respect to the number of redundant tests and the time taken per test. Here, the hunters and the scanners are those subjects who were identified as such at both levels of task difficulty. The comparison was restricted to the problems solved in the think-aloud condition as the strategy identification had been made on the basis of the verbal protocols. Furthermore, since the strategy-related behavioral differences were expected to occur in each problem, the scores on the two measures were averaged across all the (four) problems of that condition.

For the hunters (N = 8), the number of redundant tests had a mean of 0.56 and a standard deviation of 0.32. For the scanners (N = 11), the mean of this measure was 0.14 with a standard deviation of 0.23. A two-samples t-test for independent groups revealed that these means differed significantly (t(17) = 3.36, p = 0.002, one-tailed). For the hunters, the time taken per test had a mean of 77.29 and a standard deviation of 24.49. For the scanners, the mean of this measure was 108.84 with a standard deviation of 21.33. Again, a two-samples t-test for independent groups revealed that these means differed significantly (t(17) = -2.99, p = 0.004, one-tailed). Thus, on the average, the hunters needed a larger number of redundant tests and took less time per test than the scanners. This is in accordance with the *a priori* specified behavioral implications.

In conclusion, there is evidence that the data resulting from the strategy identification procedure were both reliable and valid.

3.2. Tests of primary interest

On the basis of the strategy data, the following three groups of subjects were created:

1. those being identified as scanner at both levels of task difficulty (N = 11);
2. those being identified as scanner at the low level of task difficulty and as hunter at the high level (N = 4);
3. those being identified as hunter at both levels of task difficulty (N = 8).

Adhering to the terminology used in the introduction, these groups will respectively be referred to as: the persisting subjects, the yielders and the insensitive subjects.

As described above, in both experimental conditions two problems with the same level of task difficulty were administered. Any two such problems were considered as replications and a subject's scores on these were therefore averaged before being used in the subsequent analyses.

The major statistical technique used was analysis of variance. Because of our interest in detecting measure-specific effects, it was decided to apply for each measure under study a multivariate measurement model (MANOVA). Thus, the data were not treated as multivariate with respect to scores on several measures but they were with respect to repeated scores on one measure. Whenever in a MANOVA an interaction term reached or approached significance at the 0.05 level, *post hoc* t-tests for paired samples were applied in order to disentangle the pattern of the effect.

Primary-task interference

Interference with the primary task was tested by comparing the subjects' performance in the think-aloud condition with their performance in silence.

First, the mean and standard deviation were calculated for the two primary-task measures being evaluated. This was done separately for the three strategy groups, for the two experimental conditions and for the two levels of task difficulty. Table 3.2a and b present the results of these calculations. Next, the scores on the primary-task measures were subjected to MANOVAs which included the between-subjects factor strategy group and the within-subjects factors experimental condition and task difficulty. The yielders were excluded from these analyses as they were too small in number to allow a reliable assessment of the effects of interest. For the two measures involved, a separate MANOVA was carried out. A summary of the MANOVAs is given in Table 3.3. The only concern in these analyses lay in detecting possible effects of experimental condition, i.e., either as

Table 3.2a. Mean and standard deviation of the time to completion (in seconds) calculated for the strategy groups per experimental condition and task difficulty. The mean is given at the top of a cell and the corresponding standard deviation at the bottom.

	Experimental condition			
	Think aloud		Silent	
Strategy group	Task difficulty			
	Easy	Difficult	Easy	Difficult
Persisting subjects	253.31 54.46	700.10 172.10	247.34 59.52	627.41 222.62
Yielders	506.12 210.95	638.04 244.28	416.55 277.31	551.67 270.71
Insensitive subjects	438.06 134.26	402.48 117.07	363.78 142.15	414.39 171.02

Table 3.2b. Mean and standard deviation of the number of trials calculated for the strategy groups per experimental condition and task difficulty. The mean is given at the top of a cell and the corresponding standard deviation at the bottom.

	Experimental condition			
	Think aloud		Silent	
Strategy group	Task difficulty			
	Easy	Difficult	Easy	Difficult
Persisting subjects	4.41 0.70	5.32 2.23	4.59 0.86	6.91 2.50
Yielders	4.75 0.87	6.63 0.75	5.13 1.31	7.50 4.34
Insensitive subjects	5.75 0.89	6.50 1.44	6.38 2.15	6.13 0.95

Table 3.3. Summary of the MANOVAs for the time to completion and the number of trials. A cell contains the obtained F and, in parentheses, the corresponding p. The degrees of freedom are (1,17) for each effect tested.

Performance measure	Effect						
	S	C	D	S * C	S * D	C * D	S * C * D
Time to completion (in seconds)	1.30 (0.269)	1.74 (0.204)	40.44 (0.000)	0.02 (0.880)	37.60 (0.000)	0.04 (0.846)	2.38 (0.141)
Number of trials	3.70 (0.071)	2.47 (0.134)	4.54 (0.048)	1.40 (0.253)	2.43 (0.137)	0.14 (0.718)	4.69 (0.045)

Key. S: strategy group; C: experimental condition; D: task difficulty.

main effect or in interaction with (one of) the two other factors included. The MANOVA for the time to completion showed that none of these effects reached or approached significance. The MANOVA for the number of trials, however, did show such an effect.

On this measure, the three-factor interaction proved to be significant. The nature of this interaction is clarified in Figure 3.1 which plots the means of the measure presented in Table 3.2b. In this figure, it can be seen that the insensitive subjects, whether solving easy or difficult problems, needed, on the average, about the same number of trials under thinking aloud and in silence. In this group, the difference between the two conditions did not approach significance, i.e., neither for the easy problems (t(7) = -0.79, p = 0.457, two-tailed) nor for the difficult problems (t(7) = 0.52, p = 0.621, two-tailed). The persisting subjects also exhibited no significance difference with respect to the average number trials between the two conditions, if only on the easy problems (t(10) = -1.79, p = 0.104, two-tailed). On the difficult problems, however, this group needed, on the average, a significantly smaller number of trials under thinking aloud than in silence (t(10) = -2.41, p = 0.037, two-tailed). So, when required to think aloud, the persisting subjects performed the task more accurately, at least at the high level of difficulty. It thus appears that, under certain conditions, thinking aloud changes primary-task performance.

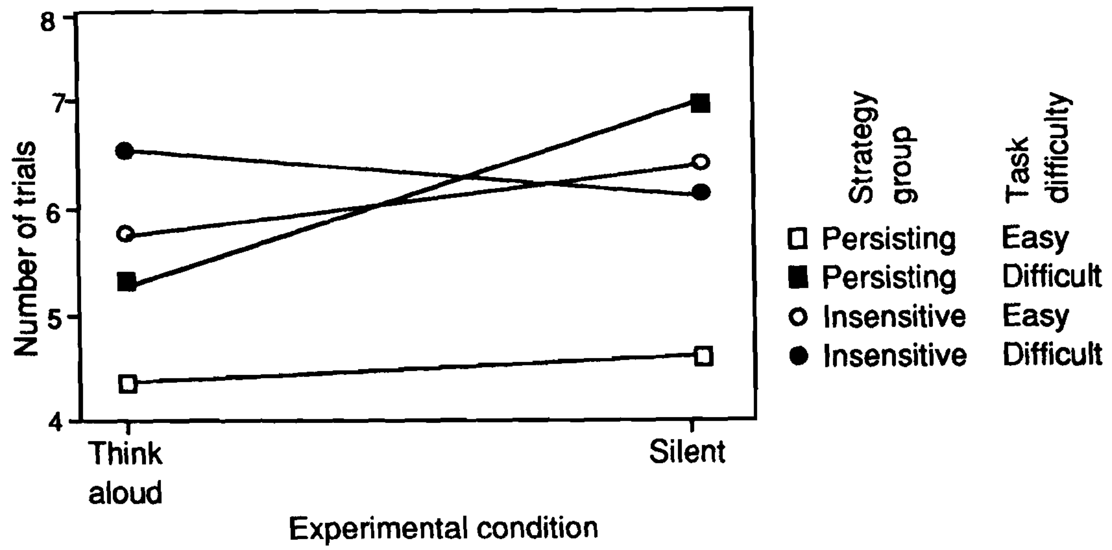


Figure 3.1. Mean number of trials for the persisting subjects and the insensitive subjects per experimental condition and task difficulty.

Secondary-task sensitivity

Sensitivity was tested by examining for the different strategy groups the effect of task difficulty on think-aloud performance.

First, the mean and standard deviation of the rate of verbalization was computed. These computations were carried out separately for the three strategy groups and for the two levels of task difficulty. Table 3.4 gives the result of this. The data on verbalization rate were further analyzed by subjecting them to a MANOVA which included the between-subjects factor strategy group and the within-subjects factor task difficulty. Because of their small number, the yielders were again omitted from the analysis. The results showed that the main effect of strategy group did not approach significance (F(1,17) = 1.15, p = 0.298) but the main effect of task difficulty did (F(1,17) = 3.20, p = 0.092). Furthermore, the interaction between this factor and strategy group reached significance (F(1,17) = 11.36, p = 0.004). The nature of the interaction is clarified in Figure 3.2 which plots the means given in Table 3.4. As can be seen in this figure, the insensitive subjects had about the same average rate of verbalization on the easy and the difficult problems. For this group, the difference between the two versions of problem did not approach significance (t(7) = -1.00, p = 0.350, two-tailed). The persisting subjects, however, verbalized on the average at a lower rate on the difficult problems than on the easy problems. This decrease appeared to be significant (t(10) = 4.09, p = 0.001, one-tailed).

Table 3.4. Mean and standard deviation of the rate of verbalization (in number of syllables per minute) calculated for the strategy groups per level of task difficulty. The mean is given at the top of a cell and the corresponding standard deviation at the bottom.

Strategy group	Task difficulty	
	Easy	Difficult
Persisting subjects	120.78 34.70	105.35 26.28
Yielders	126.85 30.31	111.33 22.42
Insensitive subjects	96.17 29.21	100.91 27.63

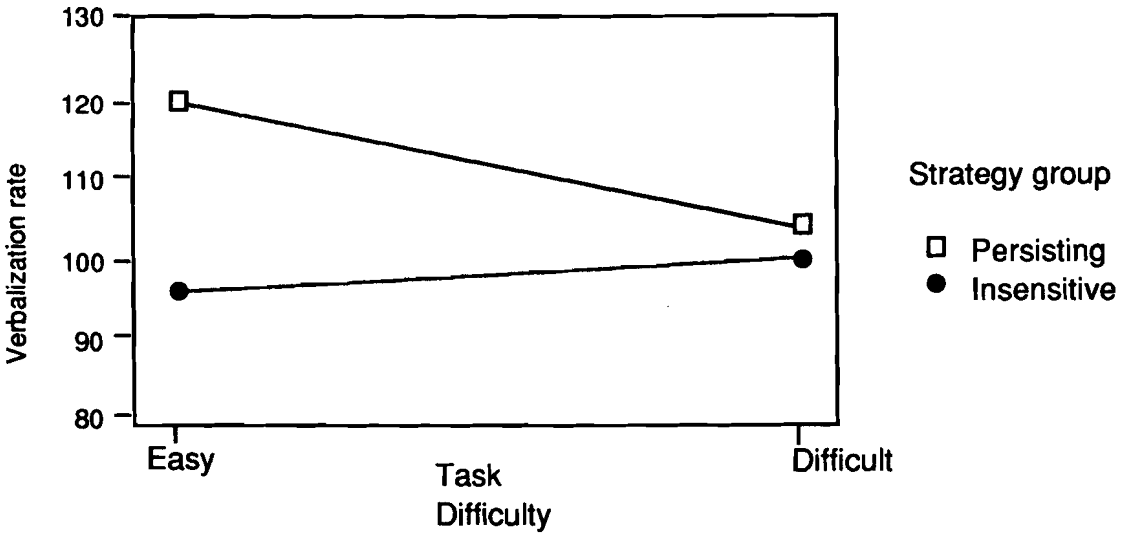


Figure 3.2. Mean verbalization rate (in number of syllables per minute) for the persisting subjects and the insensitive subjects per level of task difficulty.

4. DISCUSSION

The study shows that depending on the way a primary task is carried out increases along a dimension of difficulty may or may not produce a degradation of think-aloud performance as measured by the rate of verbalization. More specifically, the results stress the importance of distinguishing between persisting subjects (i.e., those who despite increases in task difficulty stick to their strategy, thereby facing higher cognitive demands) and insensitive subjects (i.e., those adopting a strategy for which the chosen dimension of task difficulty is not effective). As expected, the persisting subjects decreased their verbalization rate whereas the insensitive subjects did not. Unfortunately, another group of subjects, consisting of those who in response to increases in task difficulty adopted a cognitively less demanding strategy, was too small in number to allow a reliable assessment of the effects of interest. Considered in isolation, these results point to the feasibility of using thinking aloud as a secondary task for measuring the spare processing capacity associated with a primary task. However, the problem is that in one case thinking aloud interacted significantly with the primary task. In particular, the persisting subjects performed the difficult version of the primary task more accurately when verbalizing concurrently than in silence. As explained in the introduction, the occurrence of interference makes it very difficult to see secondary-task decrements as a pure index of the spare capacity of the primary task.

In the present case, the source of the interference seems to lay in the way the subjects divided their processing resources between the two tasks as a function of primary-task difficulty. To explain this point, consider our finding that the persisting subjects, when required to verbalize their thoughts, maintained performance on the easy version of the primary task but increased it on the difficult version. This finding suggests that the amount of resources the subjects allocated to the primary task under think-aloud conditions was disproportionately larger when performing the difficult version of that task than when performing the easy version. Only this extra resource allocation in favour of the difficult primary task could already have the spurious effect of degrading think-aloud performance. The question thus becomes whether the persisting subjects would also have decreased their verbalization rate if they had not changed their resource allocation policy.

Primary-task interference has frequently posed a serious problem in applications of the spare capacity task methodology. As a matter of fact, it has been argued that interference can never be eliminated completely. To deal with this problem, a number of investigators propose to use a dual-task paradigm in which cognitive load is not measured in terms of performance decrements in one of the two tasks involved but rather in terms of combined task performance decrements. This paradigm essentially consists of instructing the subjects to allocate their processing resources between the concurrently presented tasks so that their overall performance is as good as possible. Here, observed decrements on the two tasks are treated as bivariate observations in a single mutual interaction space. Such a space is usually referred to as a Performance Operating Characteristic (POC) space. In a POC space, a larger distance of a data point from the origin indicates a higher degree of interference between the tasks. As interference increases, the cognitive load sensitivity of one (or both) tasks is assumed to be better. A detailed treatment of this type of analysis, including methodological issues, can be found in Gopher and Donchin (1986) and Wickens (1984). Shingledecker and Crabtree (1982) and Wickens, Mountford, and Schreiner (1981) provide interesting applications. Note that while in the spare capacity task methodology interference is seen as a highly undesirable phenomenon, in the POC paradigm the assessment of the degree of interference actually constitutes the main objective of the measurement itself. It should also be noted that the POC methodology can be extended by varying the priorities of the two concurrently presented tasks. This involves instructing the subjects to change the way in which they divide their processing resources between the tasks from one trial to another. Since in the present experiment shifts in resource allocation policy are probably the cause of the interaction between the main task and thinking aloud, the POC methodology might be a more promising approach to evaluate the cognitive load sensitivity of verbalization-based performance measures.

Although not of primary importance to the objective of this study, there is at least one other result warranting discussion. The persisting subjects have on the average a considerably higher rate of verbalization than the insensitive subjects, although this difference is not significant and is primarily restricted to the low level of task difficulty. A possible explanation for this finding may be sought in individual differences in the ability to perform the task or in the ability to think aloud. It is conceivable that the persisting subjects are more skilful in the task or are in general more able to verbalize their thoughts than the insensitive subjects.

Another explanation is that the two basic strategies the subjects follow during the task differ in the ease of being verbalized. It could be argued that the scanning strategy adopted by the persisting subjects is more easily verbalized than the hunting strategy adopted by the insensitive subjects. Following this line of argument, however, runs counter the position taken throughout the present study. Given that the scanning strategy in comparison with the hunting strategy is more cognitively demanding, exactly the opposite result would be predicted. Here, the assumption is that the larger the demands of the primary-task strategy on the limited supply of processing resources, the less remains for concurrent verbalization. Of course, one might question the tenability of this assumption itself. It has been claimed, for example (Schneider and Shiffrin, 1977), that the amount of resources needed for cognitive processing is inversely to the speed of processing. This claim essentially means that cognitive processes requiring few or no resources proceed rapidly while cognitive processes requiring much resources have a slow speed. If this claim is valid, it might be that less demanding mental activities speed up so greatly that because of the relatively slow speech rate concurrent verbalization becomes scarce or is stopped. Contrastingly, highly demanding mental activities would then proceed slowly enough to verbalize them completely. It can thus be argued that cognitive processes which require more resources will be verbalized better rather than worse.

5. CONCLUSION

By increasing the difficulty of a primary task, concomitant decreases in think-aloud performance, as indexed by the rate of verbalization, could be observed, at least when controlling for the way the primary task was actually carried out (i.e., strategy use). However, thinking aloud interacted with the primary task so that the observed decreases in verbalization rate could not (only) be attributed to the resource expenditure associated with primary-task performance. The nature of the interaction suggested that another factor requiring consideration was the way in which the processing resources are divided between the two tasks as a function of primary-task difficulty. Therefore, the results of the study do not warrant the conclusion yet that verbalization rate can be used as a sensitive index of cognitive load. More research is needed before definite conclusions can be drawn with respect to this issue.

References

Brinkman, J.A. (1990). The analysis of fault diagnosis tasks: do verbal reports speak for themselves? Ph.D. thesis. Eindhoven, The Netherlands: Eindhoven University of Technology, Graduate School of Industrial Engineering and Management Science.

Bromme, R., and Wehner, T. (1987). Zum Zusammenhang von Sprechgeschwindigkeit und Sprechfehlern mit der Aufgabenschwierigkeit beim lauten Denken. *Zeitschrift für experimentelle und angewandte Psychologie, 34,* 1-16.

Brooke, J.B., Duncan, K.D., and Cooper, C. (1980). Interactive instruction in solving fault-finding problems: an experimental study. *International Journal of Man-Machine Studies, 12,* 217-227.

Brooke, J.B., Duncan, K.D., and Marshall, E.C. (1978). Interactive instruction in solving fault finding problems. *International Journal of Man-Machine Studies, 10,* 603-611.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Deffner, G. (1984). *Lautes Denken: Untersuchung zur Qualität eines Datenerhebungsverfahrens.* Frankfurt am Main, Germany: Peter Lang.

Deffner, G., and Ericsson, K.A. (1985). *Lautes Denken bei nicht-sprachlichen Denkprozessen: Sprechtempo und Pausen.* Paper presented at the 27th Tagung experimentell arbeitender Psychologen. Wuppertal, Germany.

Ericsson, K.A., and Simon, H.A. (1984). *Protocol Analysis: Verbal Reports as Data.* Cambridge, Massachusetts: M.I.T. Press.

Gopher, D., and Donchin, M. (1986). Workload-an examination of the concept. In: K.R. Boff, L. Kaufman, and J.P. Thomas (Eds), *Handbook of Perception and Human Performance (Vol. II): Cognitive Processes and Performance.* New York: John Wiley and Sons.

Johnson, E.S. (1978). Validation of concept-learning strategies. *Journal of Experimental Psychology: General, 107,* 237-266.

Kempkensteffen, J.(1987). Zeitliche Analyse des Sprechens beim lauten Denken. Unpublished thesis. Hamburg, Germany: University of Hamburg, Institute of Psychology II.

Knowles, W.B. (1963). Operator loading tasks. *Human Factors, 5*, 155-161.

O'Donnell, R.D., and Eggemeier, F.T. (1986). Workload assessment methodology. In: K.R. Boff, L. Kaufman, and J.P. Thomas (Eds), *Handbook of Perception and Human Performance (Vol. II): Cognitive Processes and Performance.* New York: John Wiley and Sons.

Ogden, G.D., Levine, J.M., and Eisner, E.J. (1979). Measurement of workload by secondary tasks. *Human Factors, 21*, 529-548.

Ohlsson, S. (1980). Competence and strategy in reasoning with common spatial concepts: a study of problem solving in a semantically rich domain (Report No. 6 of the Cognitive Seminar). Stockholm, Sweden: University of Stockholm, Department of Psychology.

Rhenius, D., and Deffner, G. (1990). Evaluation of concurrent thinking aloud using eye-tracking data. *Proceedings of the Human Factors Society 34th Annual Meeting.* Orlando, Florida, 1265-1269.

Russo, J.E., Johnson, E.J., and Stephens, D.L. (1989). The validity of verbal protocols. *Memory and Cognition, 17*, 759-769.

Schneider, W., and Shiffrin, R.M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84*, 1-66.

Shingledecker, C.A., and Crabtree, M.S. (1982). Subsidiary radio communications tasks for workload assessment in R&D simulations: II. Task sensitivity evaluation (Report No. AFAMRL-TR-82-57). Wright-Patterson Air Force Base, Ohio: U.S. Air Force Aerospace Medical Research Laboratory.

Simon, D.P., and Simon, H.A. (1978). Individual differences in solving physics problems. In: R.S. Siegel (Ed.), *Children's Thinking: what develops?* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Sperandio, J.C. (1978). The regulation of working methods as a function of workload among air traffic controllers. *Ergonomics, 21*, 193-202.

Wickens, C.D. (1984). *Engineering Psychology and Human Performance.* Columbus, Ohio: C.E. Merrill Publishing Company.

Wickens, C.D., Mountford, S.J., and Schreiner, W. (1981). Multiple resources, task-hemispheric integrity, and individual differences in time-sharing. *Human Factors*, *23*, 211-229.

Williges, R.C., and Wierwille, W.W. (1979). Behavioral measures of aircrew mental workload. *Human Factors*, *21*, 549-574.

Eindhoven University of Technology
Department of Industrial Engineering and Management Science
Research Reports (EUT-Reports)

The following EUT-Reports can be obtained by writing to:
Eindhoven University of Technology, Library of Industrial Engineering and Management Science, Postbox 513, 5600 MB Eindhoven, Netherlands.
The costs are HFL 5.00 per delivery plus HFL 15.00 per EUT-Report, to be prepaid by a Eurocheque, or a giro-payment-card, or a transfer to bank account number 52.82.11.781 of Eindhoven University of Technology with reference to "Bibl.Bdk", or in cash at the counter in the Faculty Library.

20 LATEST EUT-REPORTS

EUT/BDK/57	Trends and tasks in control rooms **T.W. van der Schaaf**
EUT/BDK/56	The system of manufacturing: A prospective study **J.C. Wortmann, J. Browne, P.J. Sackett**
EUT/BDK/55	Rekenmodellen voor de grootschalige mestverwerking; gebaseerd op het MEMON-mestverwerkingsprocédé **Mat L.M. Stoop**
EUT/BDK/54	Computer, manager, organisatie (deel I en II) **R. Cullen, H. Grünwald, J.C. Wortmann**
EUT/BDK/53	Risico diagnose methode voor produktinnovatieprojecten; Een uitwerking toegesneden op de Industriegroep TV van Philips Glas te Eindhoven/Aken **J.I.M. Halman, J.A. Keizer**
EUT/BDK/52	Methodological problems when determining verbal protocol accuracy empirically **J.A. Brinkman**
EUT/BDK/51	Verbal protocol accuracy in fault diagnosis **J.A. Brinkman**
EUT/BDK/50	Techniek en marketing **H.W.C. van der Hart**
EUT/BDK/49	Een methoden voor kosten-batenanalyse voor automatiseringsprojecten bij de overheid **M. van Genuchten, F. Heemstra, R. Kusters**
EUT/BDK/48	Innoveren in technologie-gedreven ondernemingen, bedrijfskundige aspekten van de voorontwikkelfunktie **W.H. Boersma**
EUT/BDK/47	The creation of a research model for estimation **M. Howard**
EUT/BDK/46	Het 80 flat square project; Een case studie als aangrijpingspunt voor lerend innoveren **J.I.M. Halman, J.A. Keizer**
EUT/BDK/45	Interface design for process control tasks **T.W. van der Schaaf**
EUT/BDK/44	Afzetfinanciering **S.G. Santema**
EUT/BDK/43	Het gebruik van natte (industriële) bijproducten in de varkenshouderij; Een verkenning van de Nederlandse situatie **Mat. L.M. Stoop**
EUT/BDK/42	An integral approach to safety management **T.W. van der Schaaf**
EUT/BDK/41	De produktie van varkensvlees; Een integrale ketenbenadering Deelrapport 1: Enkele modellen voor de varkenshouderij **A.J.D. Lambert**
EUT/BDK/40	Informatievoorziening ten behoeve van klantenorder-acceptatie; een eerste verkenning **F.J. Faszbender**
EUT/BDK/39	A bibliography of the classical sociotechnical systems paradigm **F.M. van Eijnatten**
EUT/BDK/38	Meten van kwaliteit van Nederlandse instrumentatie op basis van ontwerpgerichte toepassingsaspekten **F.M. van Eijnatten**