Methodological problems when determining verbal protocol accuracy empirically

Citation for published version (APA):

Document status and date:
Published: 01/01/1992

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Methodological problems when determining verbal protocol accuracy empirically

by J.A. Brinkman

Report EUT/BDK/52
ISBN 90-386-0171-9

Eindhoven 1992
Brinkman, J.A.

Methodological problems when determining verbal protocol accuracy empirically / by J.A. Brinkman
Eindhoven: Eindhoven University of Technology.
Graduate School of Industrial Engineering and Management Science.
(Report EUT/BDK;52)
With references.
ISBN 90-386-0171-9
Subject headings: verbal protocol analysis.
Methodological problems when determining verbal protocol accuracy empirically

Jan A. Brinkman

Abstract

The approach normally followed to provide an empirical assessment of verbal protocol accuracy is evaluated from a methodological point of view. Three classes of problems are identified with this approach. These pertain to: selecting and eliciting recordable behaviors, converting the observations into data and interpreting the data. Possible ways to deal with the listed problems are presented. It is emphasized that, when adopting the approach, verbal protocol accuracy can only be inferred rather than directly observed.

1. Introduction

At present, the verbal protocol methodology is widely applied as a research tool for analyzing the cognitive processes that are used while performing a particular task. The application of this methodology is based on the idea that a subject's verbalizations about what he is or has been thinking during task performance is as close as one can get to his real cognitive processes. At the same time, it is generally recognized that the verbal protocol methodology is associated with a number of problems which can affect the quality of the data it yields. First, there is the question whether the verbal reports that are obtained suffer from reactivity. Reactivity occurs if the cognitive processes which normally proceed during task performance are changed by the mere requirement to verbalize them concurrently or retrospectively. For example, when there are several ways to perform a task, it may be that verbalization has the effect that the normal way is changed for a way which is more easily described. Second, there is question whether the verbal reports are really valid. When lacking validity, the reports do not adequately capture the cognitive processes that actually going on, irrespective of the question whether the processing is reactive or not. For example, in a highly practised task, the speed of the ongoing cognitive processes may proceed the ability to verbalize them completely.

Reviewing the relevant literature, it is clear that the most prominent theoretical model of verbal reporting currently available has been developed by Ericsson and Simon (1984). Unfortunately, there appear to be several instances where the predictions drawn from this model do not come true (see e.g. Russo et al., 1989). If it is impossible to predict verbal report accuracy on the basis of theoretical notions, attempts should be undertaken to determine it empirically. It is the purpose of this
paper to provide a critical evaluation of the approach which is normally followed to accomplish this. More specifically, the paper discusses a number of methodological problems of the usual approach of testing verbal report accuracy empirically.

2. The usual approach for an empirical assessment of verbal protocol accuracy

Basically, the approach in question consists of adding a silent control condition to the condition in which the subjects have to verbalize their thoughts (see Fig. 1).

<table>
<thead>
<tr>
<th>silent control condition</th>
<th>verbalization condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>performance data</td>
<td>performance data</td>
</tr>
<tr>
<td>protocol data</td>
<td></td>
</tr>
</tbody>
</table>

| reactivity | validity |

Figure 1. The usual approach for determining verbal protocol accuracy empirically.

In both conditions, a set of performance data is derived from the non-verbal behavioral recordings that are made. For example, the strategies the subjects follow
during task performance might be identified by tracking their eye movements. Then, the two sets of data obtained are compared. The better the agreement, the less reactive the verbal reports are assumed to be. In addition, in the verbalization condition a set of protocol data is extracted from the recorded verbalizations. This set is compared with the set of performance data which are collected in the same condition. The better the agreement here, the more valid the verbal reporting is assumed to be. For a nice demonstration of the application of this research methodology, see Deffner (1984) or Rhenius and Deffner (1990).

In the approach presented so far, it has been supposed that the determination of verbal report validity is restricted to one particular verbalization condition. However, it is possible to extend the approach in this respect by introducing a carefully selected experimental factor, e.g. the nature of the task, of which the operationalization results in several verbalization conditions. Then, verbal report validity can be determined by establishing whether the created verbalization conditions result in performance differences which correspond with variations in the protocol data (see e.g. Berry and Broadbent, 1984).

In the literature, an increasing number of studies is reported which have explicitly been designed to provide an empirical assessment of verbal protocol accuracy. In practically all these studies, the researchers follow an approach which can relatively easily be fitted in the approach described here. It appears, however, that many researchers test only one of the two forms of verbal report accuracy, i.e. either non-reactivity or validity, rather than both. Furthermore, in much research it is common practice to base the test on data which capture only the outcomes, the products, of the cognitive processes engaged in, such as total time on task and overall accuracy. Only a few researchers use data which represent the cognitive
processes themselves, the underlying operations, like the strategy or the heuristics employed. In the following section, the implications of these restrictions will become clear.

3. Methodological problems with the approach

Assessing verbal protocol accuracy according to the way it is generally done is not so simple as it seems. Indeed, the apparent simplicity of this approach might divert attention from the methodological problems to be encountered when trying to arrive at sound conclusions. These problems may be classified according to the specific stages in which one is successively involved when following the approach, namely:

1. selecting and eliciting observable behaviors;
2. converting the observations into data;
3. interpreting the data.

In this section, we will be concerned with these three classes of problems and, where possible, with the different ways to handle them.

3.1. Selecting and eliciting observable behaviors

3.1.1. Non-verbal behaviors

One of the problems associated with the approach is that it can only be applied for the study of tasks where the cognitive processes involved underlie non-verbal behaviors which are suitable for being recorded. Non-verbal behavioral recordings are
typically needed for generating the required performance data. However, in many
task settings, the cognitive processes have no or hardly any counterparts in
recordable non-verbal activities. Consequently, the research methods which are in
principle available to carry out these recordings are usually not robust enough to be
used in the variety of settings in which the verbal protocol methodology is
applicable. For instance, the method of tracking eye movements is restricted to tasks
which have strong visual components. Furthermore, the measurements provided by
these methods are generally too coarse given the great deal of detail provided by the
corresponding verbalizations. For example, the temporal density of the measurements
obtained when recording the manipulative responses of a subject during a problem-
solving task may be much lower than the density of the stream of verbalizations this
subject produces concurrently.

3.1.2. Verbal behaviors

Distinguishing between non-reactivity and validity does not mean that the
interdependency between these two forms of verbal report accuracy should be
neglected. In fact, the likelihood of generating a verbal report which is accurate in
one sense depends upon the likelihood of generating it accurately in the other. This
can be seen as follows (free after Russo et al., 1989).

When producing a verbal report, two kinds of information-processing activities
have to be carried out: those directed at the task to be done and those related to the
act of verbalization. Adequate performance of each of these two activities requires a
certain amount of attentional capacity. However, on the assumption that this capacity
is not unlimited, difficulties may arise. That is to say, it may be impossible to divide
the available attentional capacity between the two types of activity so that both are
adequately performed. If the task-directed processes receive less attention, they will be more easily affected by the act of verbalization. Then, it becomes more likely that the resulting verbal report suffers from reactivity. If, on the other hand, less attention is paid to verbalizing, the task-directed processes will be less well captured as a result of which the likelihood of an invalid verbal report increases.

Attentional capacity withdrawn from one type of activity can in principle be utilized for the other type so that one form of verbal report accuracy may be exchanged for the other form. Such a trade-off can be operative in any study designed to test verbal report accuracy and how it will manifest itself is dependent upon the procedures that are adopted to elicit the verbalizations. These procedures include instructions, training, conditions of testing, and so on.

Consider, for example, a study in which a number of subjects verbalize their thoughts while performing a particular task. Suppose first that in this study the conditions are arranged so that whenever the subjects remain silent for more than five seconds, they are reminded to resume verbalizing. In their endeavours to obey these reminders, the subjects might change the cognitive processes normally engaged in. As a result, the verbal reports produced by them are valid but suffer from reactivity. Suppose now that it is decided to refrain from using verbalization reminders. Then, the subjects will not feel obliged to verbalize accurately and are thus able to preserve their normal cognitive processes. Consequently, the resulting verbal reports are not reactive but invalid.

This example makes clear that if the test procedures put the emphasis on one of the two forms of verbal report accuracy, the other form may not be realized. So, when testing verbal report accuracy, one is faced with the problem whether one or the other form, if not both, should be emphasized.
Apart from the possibility that the two forms of verbal report accuracy trade off, it is also conceivable that inaccuracy in one form is simply caused by inaccuracy in the other. For example, under the influence of a verbalization request, a subject may change the strategy he normally adopts, but by doing so he may no longer be able to report accurately about the cognitive processes now engaged in. He thus produces a verbal report which is reactive as well as invalid.

How should dependencies between the two forms of verbal report accuracy be dealt with? There is no simple solution to this problem and the best way to handle it is eventually determined by the exact goal of the study being carried out. With respect to this problem, it is also relevant to differentiate the situation in which both forms of verbal report accuracy are tested from the situation in which only one form is tested.

**Testing both forms of verbal report accuracy.** In practice, verbal reports are typically gathered to gain insight into the cognitive processes as these normally proceed. It is therefore of paramount importance that the very same processes occur when reporting about them. Only after this has been established does it become essential that the processes are accurately reported. In other words, it is not until the non-reactivity of a verbal report has been ascertained that its validity is called into play. Indeed, when not fully confident whether a particular verbal report is free from reactivity, one would do better to refrain from using it immediately. Otherwise, one runs the risk of drawing incorrect conclusions about the cognitive processes supposedly being reported. Of course, as soon as a verbal report has been given up because of its assumed reactivity, there is no need to be worried about its validity any longer. Seen from this practical point of view, it follows that, when examining both forms of verbal report accuracy, one should in the first place take care that the reporting is not
affected by reactivity and only then try to keep it free from invalidity. Possible ways to realize this are suggested by Ericsson and Simon's model of verbalization (1984). According to this model, reactivity is, for instance, reduced if the verbalization instructions ask the subjects to report about the thought processes as these actually occur during task performance rather than asking them to provide information which has to be inferred, such as explanations for their thoughts.

Before proceeding, it is necessary to qualify the preceding passage somewhat. First, several forms of reactivity and invalidity can be distinguished and a verbal report may suffer from one form but not from another (Russo et al., 1989). Which form should be avoided and which one may be allowed ultimately depends upon the purpose for which the verbal report is used. Suppose, for example, that a number of concurrent verbal reports is collected in order to get a better understanding of the course and structure of the cognitive processes involved in a problem-solving task. Then, it may be taken for granted that concurrent verbalization leads to a slowing down of these processes as long as they are not changed fundamentally. Second, the reactivity or invalidity of a verbal report is not all-or-nothing but rather a matter of degree. It is, for instance, possible that a verbal report shows many gaps and still contains a sufficient number of verbalizations to use it for the intended purpose.

Testing only one form of verbal report accuracy. Although in order to be useful a verbal report should be non-reactive as well as valid, it may happen that for one reason or another one decides to test only one of these two forms of accuracy. Should in this case an accurate verbal report result, it can still be seen to it that a more or less clear picture of its usefulness is obtained. This is accomplished by designing the test situation so that the verbal report satisfies at any rate the accuracy form which is not being tested. The more care taken that the verbal report is accurate
with regard to the not-tested form, the greater the certainty that the report is useful, provided that it is accurate with respect to the tested form. Thus, whenever only one form of verbal report accuracy is examined, it is, at least from a practical point of view, better to use all the available opportunities to achieve the other, not-tested form instead of leaving this one uncontrollable. Once again, Ericsson and Simon's model of verbalization (1984) may be used as a starting point to realize this.

Testing only one of the two forms of verbal report accuracy and at the same time trying to realize the other form, is in general a less strong approach than performing a test on both. The reason for this is that, when adopting the former approach, one may be in doubt whether the not-tested form has really been achieved. What can be done, however, is to collect circumstantial evidence on this point. If, for instance, a verbal report is subjected to a reactivity test, one might still get some indication of how well this report fulfills the validity requirement by determining whether it contains a sufficient number of verbalizations for a content analysis. But of course, this sort of evidence does not possess the conclusive power of a validity test in which the protocol data on hand are directly related to the corresponding performance data.

3.2. Converting the observations into data

Another problem associated with the approach under consideration is that application of it is only warranted if a number of relatively restrictive assumptions about the collected data has been satisfied. Furthermore, as will become clear soon, verification of these assumptions is in itself no sinecure. The strongest assumptions being made pertain to:
1. the validity of the performance data (see e.g. Deffner, 1984 or Russo et al., 1989);
2. the reliability of the protocol data (e.g. Ericsson and Simon, 1984);
3. the sensitivity of these two types of data (e.g. Berry and Broadbent, 1984).

Each of these assumptions will be considered in turn below.

3.2.1. Validity of performance data

In the approach, the performance data are essentially employed as the criterion for making decisions about verbal report accuracy. Specifically, the performance data constitute the basis for determining the reactivity of verbal reporting and these data are also the basis against which the protocol data are judged when determining verbal report validity. Using performance data for suchlike criterion purposes rests on the assumption that these data are perfectly valid, i.e. give an accurate picture of the cognitive processes engaged in. Note that this assumption is no triviality. As a matter of fact, when testing verbal report accuracy, it is not sufficient to establish that the two sets of data which are compared agree with one another. This is because it is in principle possible to achieve perfect agreement while both sets of data suffer from invalidity.

Of course, the point is whether the assumption of valid performance data is always tenable. With respect to this point, it is useful to refer to the already made distinction between product-oriented data and operation-oriented data. In the approach, the performance data are obtained by selecting non-verbal behaviors for measurement and by encoding the observations made. If the performance data are
product-oriented, this encoding process proceeds in a more or less straightforward way. This is because non-verbal behaviors typically reveal (a portion of) the outcomes of the underlying cognitive processes. As a consequence, product-oriented performance data can be qualified as inherently valid. On the other hand, the validity of operation-oriented performance data may be less evident as the cognitive processes themselves are in general poorly reflected in the accompanying non-verbal behaviors.

So, when deriving operation-oriented performance data, one must usually rely on strong theoretical presuppositions, i.e. a model of the cognitive processes used in the task being studied. In this case, it is the accuracy of the cognitive task model invoked which determines the validity of the performance data. Then, there is clearly a need to provide some evidence that the collected data are really valid.

An appropriate way to demonstrate the validity of a set of operation-oriented performance data is to show that they maintain previously specified relationships with other behavioral measures (Johnson, 1978). It is here where the cognitive task model might play a role again as it could be helpful in predicting these relationships. Of course, this demonstration is only convincing if the other measures are initially not used in the derivation of the performance data themselves. For example, in a given problem-solving task it may be that according to the associated cognitive model, the use of one particular strategy entails a greater amount of information processing than any other strategy. Then, it seems reasonable to expect that users of this strategy will take more time to complete the task than users of other strategies. Confirmation of this prediction would increase confidence in the validity of the identified strategies, given that time to task completion is not a variable directly involved in the original strategy identification. However, although such a demonstration may be regarded as circumstantial evidence, it does certainly not mean
that in this way the validity of the performance data is really proven. At best, the results will only show no (marked) violations of the assumption.

### 3.2.2. Reliability of protocol data

The validity of verbal reporting is tested by encoding the recorded verbalizations and validating the resulting protocol data. In doing so, it is assumed that the encoding proceeds reliably, that is to say in such a manner that the protocol data are reproducible. Actually, the protocol data cannot be valid unless they are reliable. It is thus essential that the reliability of the protocol data is ascertained before their validity is tested. This can be done by determining inter- and intra-coder agreement.

Inter-coder agreement relates to the extent to which the same protocol data are produced by a number of different coders. This form of agreement is tested by determining how well two or more coders extract independently of each other similar encodings from the same verbal report. In a similar vein, intra-coder agreement refers to the extent to which a single coder is able to reproduce protocol data over time. This form of agreement is tested by establishing how well one coder extracts similar encodings when analyzing the same verbal report on two or more occasions.

The reliability of the protocol data depends upon the encoding process according to which these data are generated. This encoding process usually consists of having one or more coders classify the verbal reports on hand into a number of a priori defined categories which are derived from a cognitive model of the task analyzed. Several factors seem to determine whether such a set of categories will be reliably applied by the coders involved. Although these factors have already extensively been discussed by Ericsson and Simon (1984), it might be worthwhile to repeat their discussion briefly and to raise a number of additional issues.
One factor playing an important role is the exactness of the definitions of the selected categories. Of course, clear and strict category definitions will help increase the ease and thus the reliability with which the categories are applied. However, an appropriate set of clearly defined categories is not always available in advance (see e.g. Rasmussen and Jensen, 1974). If this is the case, the categories may be iteratively developed from the verbal reports themselves. But then it is desirable that the ultimate categories are applied by coders who were not involved in the category development. This will still permit an adequate reliability test of the resulting protocol data (Bainbridge, 1979).

Another crucial factor is the nature of the categories. Dependent upon the goal of the study, one might decide to select a set of product-oriented categories or a set of operation-oriented categories. When using product-oriented categories, it is possible to encode the available verbal reports in a more or less direct way. According to Ericsson and Simon, this is because a verbal report (partly) reflects the information under focal attention which primarily consists of the outcomes of the cognitive processes being reported about. Consequently, encoding a verbal report into product-oriented categories can in general reliably be done. In contrast to this, a verbal report appears to be less suited for encoding it reliably into operation-oriented categories. The operations of cognitive processes are not attended to and thus remain unreportable. Therefore, they cannot be encoded directly from a report but must be inferred from it. It is for this reason that Ericsson and Simon advocate the use of product-oriented categories, at least in the first stage of the protocol analysis. Only after the verbalizations have been described in product-oriented terms, should proposals for the underlying operations be introduced to account for these data.
A final factor of importance is the level of analysis at which the encoding into the selected categories is carried out. In the case of high-level encoding, the verbal report is encoded in its entirety. So, a coding decision has to be made about the complete report. In low-level encoding, the verbal report is first broken down into a number of segments that represent meaningful stages in the processes reported about. Upon such a segmentation, the individual segments are successively encoded. Care is taken that the encoding of each segment proceeds independently of the other segments. That is to say, one tries to reduce the possibility that the coder, while analyzing the verbalizations contained in the current segment, relies on the verbalizations from the preceding or subsequent segments. The reason for doing this lies in obtaining as much independent evidence as possible in favor of or against verbal report accuracy.

When involved in high-level encoding, the coder has to consider all the verbalizations making up the verbal report simultaneously. In addition, he must integrate all these verbalizations into one single encoding decision. Such requirements are not imposed on him, or at least to a lesser degree, in the case of low-level encoding. It could thus be argued that low-level encoding is easier and therefore more reliable than high-level encoding.

This, however, does not mean that low-level encoding is wholly unproblematic. One problem is that the required segmentation can result in one or more segments which by themselves do not contain a sufficient number of verbalizations for arriving at accurate encoding decisions. This problem arises when the verbalizations referring to the constructs to be encoded, for instance strategies, are not restricted to one segment but are scattered throughout the report. Thus, for certain encoding decisions, it may be desirable that the coder, while analyzing the current segment, gets the
opportunity to consult the preceding or the next segments. However, by allowing this, there no longer exists strict independence among the encodings of the different segments. So, while in this way the reliability of the encodings may be improved, their independence is reduced. A solution to this has been proposed by Ericsson and Simon which consists of making the segments so large that each one contains just sufficient information for making the required encoding decisions. Of course, this may imply that, in certain cases, the whole protocol constitutes one segment. Obviously, there is a trade-off between the fineness of the level of analysis and the reliability to be achieved.

Given that verbal protocol analysis may take considerable time and effort, it is not surprising that many investigators have developed all kinds of tools for facilitating the encoding process. Although it may be desirable to automate the encoding process completely or partially, it seems that this is only feasible for task domains that have a reasonable amount of structure. For less well-structured task domains, interactive computer programs have been designed where the human coder is required to make the encoding decisions (e.g. Sanderson et al., 1989a).

3.2.3. Sensitivity of performance data and protocol data

As has been discussed in the previous section, testing verbal report accuracy basically consists of comparing two sets of data. The aim of this comparison is to determine whether an observed difference between the two data sets is large enough to decide that the verbal reporting is inaccurate. However, such a decision can only be made if the data are sensitive enough to detect true deviations from verbal report accuracy. Furthermore, one should be concerned that the data to be compared may have a differential sensitivity, a problem which becomes especially prominent if
validity is tested since in that case the two sets of data have a different behavioral origin.

There are several ways to accomplish an increase in data sensitivity. One way is to base the verbal report accuracy test on data of which the orientation has been attuned to the task situation under study. Let us discuss this point in more detail.

Consider a reactivity test involving a comparison between a verbalization condition and a silent control on the basis of product-oriented data. If the two conditions differ on these measures, the produced verbal reports are said to be reactive. It may be that the observed differences result from changes in the underlying operations which in turn are brought about by the act of verbalization. If, however, no reliable differences are found it is not justified to state that the verbal reports are not reactive. In fact, a reactivity effect may be operating in the sense that verbalization changes the cognitive operations without affecting the products thereof. In other words, it is conceivable that, under the influence of verbalization, the outcomes of the cognitive processes remain unaltered whereas the processes themselves undergo significant changes. Suppose, for instance, that verbalizing the cognitive processes engaged in while performing a problem-solving task provokes a less time-consuming strategy. Then, the slowing down effect brought about by the act of verbalization could be neutralized as a result of which the solution is found in the same time normally required. Thus, by examining only the outcomes of the cognitive processes being verbalized, one runs the risk of overlooking a true reactivity effect which might have been detected would the processes themselves have been considered.

That such a risk should be taken seriously is clearly demonstrated in a study carried out by Biehal and Chakravarti (1989). Working in the field of consumer
behavior, these investigators found that, under certain conditions, concurrent verbalization affected some aspects of choice processing while the effects did not involve choice outcomes. Of course, the reverse may also occur: a change in the products of cognitive processing without a fundamental change in the processing itself. For example, it is possible that a request to verbalize concurrently leads to a longer time on task but has no strong effect on the adopted strategy (see e.g. Deffner, 1984, experiment 1). In this respect, it is also important to point out that, according to Deffner (1989), the general failure to control for the operations of cognitive functioning might explain why the studies on the reactivity of concurrent verbal reporting show a disappointing lack of consistency in results.

To conclude, a more sensitive test of verbal report reactivity may be obtained by using product- as well as operation-oriented data instead of one of these two types of data alone. As will be evident, this conclusion and the preceding argument also hold when the validity of verbal reporting is tested. And indeed, Sanderson (1990a) has recently provided suggestive evidence that the conflicting verbal report validity results found with Berry and Broadbent's dynamic system control tasks (1984) may be reconciled by analyzing the strategies the subjects follow in addition to analyzing their overall performance.

Another way to increase the sensitivity of the data to be collected is to take care that they do not suffer from random variations which might obscure verbal report inaccuracy. Of course, the data have a behavioral origin and will therefore always vary randomly to a certain extent. This makes them inherently insensitive. Nevertheless, it is often possible to reduce some of the random variation in the data. The possibilities at issue here are essentially the same as those occurring in any experiment in which the effect of some independent variable is measured against the
random variation in the dependent variable. One possibility is to increase the number of observations from which the data are derived. This can be achieved by having the same subject produce several verbal reports or by gathering the verbal reports from a larger number of different subjects. Then, the average or median observation can be taken so that the data will be less distorted by extreme observations which otherwise would substantially increase random variation. Or, to give another example, when the verbal report accuracy test encompasses two or more experimental conditions, it is possible to have each subject produce a verbal report under all conditions instead of having different subjects produce a verbal report in each condition. The rationale for doing this is that repeated observations on a single subject generally result in less random variation in the data than do observations taken from different subjects (see e.g. Winer, 1971).

An appropriate way to examine whether the data used for determining verbal report accuracy are sensitive enough for this consists of manipulating some experimental factor which is expected to exert significant behavioral effects. Once again, these predictions may be drawn from a cognitive model of the task. For example, one may anticipate that the collected data will reveal a steady performance degradation with an increase in task difficulty. If the data fail to show the expected effects and additionally do not indicate any inaccuracy in the verbal reports obtained, then one will be in doubt as to the sensitivity of these data. If, on the other hand, the manipulated factor does produce the expected effects, then one can have more faith in the data's sensitivity. Note, however, that this experimental method does not prove data sensitivity. This is because it is always possible that the data are indeed susceptible to the manipulated factor but not to true verbal report inaccuracy.
From the foregoing it should be clear by now that an empirical assessment of verbal report accuracy according to the usual approach in the ideal case requires an adequate model of the cognitive processes used in the task under investigation. Such a model is in particular needed for (1) deriving the required data from the behavioral recordings, and (2) determining whether the data fulfill the underlying assumptions.

In psychological research, a great deal of model building has been done for relatively artificial and well-structured tasks, but more realistic and complex task domains are seldom covered. For exceptions, see the work of Rogalski and Samurçay (1989), Sanderson (1989; 1990b), Sanderson et al. (1989b), and Sundström (1989). Having no appropriate cognitive model for the specific task analyzed, a necessary and first step to be able to test verbal report accuracy effectively is to construct one. Apart from the fact that developing a cognitive task model may take considerable time and effort, a number of more fundamental difficulties may arise. First, on what kind of data should the cognitive task model be based? Preferably, use should be made of directly observable behavioral characteristics, such as response latencies and eye fixations, but in certain task settings it does not make sense to collect such data. Then, it may be that one is driven back on verbal reports, the accuracy of which is just under debate. Second, how should the constructed cognitive model be validated? In the ideal case, validation of a cognitive model consists of deriving a series of model-based predictions which are subsequently experimentally verified. A first step towards this is made when checking the assumptions of the verbal report accuracy test. Data which indicate that these assumptions have been met give also credit to the cognitive model. Nevertheless, much more experimentation may be needed to ascertain that the model is really adequate in the form it is used.
3.3. Interpreting the data

As repeatedly stated, the usual approach of testing verbal report accuracy consists of determining how well one set of data tends to agree with another set. Here, the assumption is that the better the agreement between the two data sets, the more accurate the verbal reporting is. This, however, should not be interpreted to imply that disagreements between the two sets necessarily result from an inability to produce an accurate report. There are a good many other factors which may be responsible for this (see also Haider-Hasebrink, 1990; Marescaux and Karnas, 1991; Sanderson, 1989 and 1990a; White, 1988). Two of these follow more or less directly from the foregoing discussion: (1) testing only one form of verbal report accuracy without controlling for the other form, and (2) basing the test on data which violate one or more of its underlying assumptions. There are, however, at least two other factors to be identified, namely (3) adopting inadequate verbalization procedures, and (4) comparing two sets of data which are partly or completely independent of one another. The latter factor may require some additional explanation.

Especially in a validity test it is very well possible that the two sets of data being compared do not agree with each other in spite of the fact that both sets are perfectly valid. This may occur when the performance data capture components of cognitive functioning which are not captured by the protocol data, or vice versa. In other words, both sets of measures may provide an accurate picture of the cognitive processes supposedly referred to, but the particular aspects tapped by one set may be so different from those tapped by the other that the two sets disagree. Remember that in deriving the performance data use is made of non-verbal behavioral recordings which, for instance, consist of the pieces of information a subject sequentially selects
while carrying out the task to be done. The protocol data, on the other hand, are extracted from the recorded verbalizations which may reveal why the subject selects a particular piece of information and how he evaluates it. Thus, whereas in this case the performance data encompass the actual actions of the subject, the protocol data incorporate his planning and evaluative responses. Of course, the subject's actions need not be consistent with his plans or evaluations. For example, when planning which piece of information to select next, the subject might adopt a demanding strategy but, while doing so, he might make a mistake and accidentally select a piece of information which is indicative of a less demanding strategy. Whenever such inconsistencies occur, the performance data and the protocol data will tend to disagree. Then, observed disagreements between the two sets of measures may be due to the fact that different aspects of the task-related processes are captured rather than being due to true verbal report inaccuracies.

To conclude, before disagreements between the collected data can be taken as evidence against verbal report accuracy, all kinds of factors must be ruled out or accounted for in some way. So, one should be very careful in deciding all too quickly that accurate verbal reporting cannot possibly be achieved.

In the foregoing, it has already been discussed what steps have to be undertaken in order to eliminate or to evaluate possible effects of the first two factors listed above. For example, detecting possible departures from the test assumptions being made would ideally require a cognitive model of the task being investigated. But how should one control for the two other factors mentioned here? For example, on what grounds is it reasonable to assume that in a particular validity test the performance measures tap the same aspects of cognitive processing as the corresponding protocol measures? As a matter of fact, if there are no strong reasons
for expecting such a correspondence beforehand, it makes little sense to perform the test at all. Obviously, one does not only need a model of the task-directed processes but also a model specifying whether, and if so, under what conditions these processes can be reported (Bainbridge, 1979). In other words, when testing verbal report accuracy empirically, one has to understand it theoretically. Given that up till now a fully adequate theory of verbal reporting is lacking, this requirement may not be met so easily.

4. Epilogue

The approach which is normally followed to provide an empirical assessment of the reactivity and validity of verbal reporting is associated with a large number of methodological problems. This, however, should not leave the reader with the idea that this research tool should no longer be relied upon for obtaining valid indications of verbal report accuracy. But, what has to be stressed is that the approach does not allow to prove verbal report accuracy with certainty. This is because the conclusions drawn will always depend upon not-directly verifiable presuppositions about the cognitive processes involved. Therefore, when following the approach, one should be acquainted with the methodological problems to be encountered and look for possible ways to deal with these.

References


EINDHOVEN UNIVERSITY OF TECHNOLOGY
GRADUATE SCHOOL OF INDUSTRIAL ENGINEERING AND MANAGEMENT SCIENCE
RESEARCH REPORTS (EUT-Reports).

EUT-reports can be obtained, as long as stock permits, by writing to
Eindhoven University of Technology, Library of Industrial Engineering
and Management Science, P.O. Box 513, 5600 MB Eindhoven, Netherlands.
The cost per delivery are HFL 3,50 plus HFL 1,50 per EUT-report.
Only payments by Eurocheque will be accepted.

20 LATEST EUT-REPORTS

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 aspecten van de voorontwikkelfunctie 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.J.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
M. van Genuchten, F. Heemstra, R. Kusters
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 toepassingsaspecten F.M. van Eijnatten
EUT/BDK/37 De toepassing van vaardigheden bij de specificatie van het
bewerkingvoorschrift D.R. Muntslag
EUT/BDK/36 Selection of Software Cost Estimation Packages
F.J. Heemstra, M.J.M. van Genuchten, R.J. Kusters
EUT/BDK/35 Zoekboek Arbeidssystemstructurering: een overzicht van
criteria voor autonomie groepen P.J.M. Berger,
R.E.F. van den Heuvel, M.H.M. Rietrae, P.C.M. Simons, onder
redactie van F.M. van Eijnatten
EUT/BDK/34 Organisatie van produktinnovatieprocessen in middelgrote
ondernemingen; een verslag van zes case-studies in de
kunststofindustrie H.C. van der Hek-de Keyser,
C.C. Krijger
EUT/BDK/33 Innovatie gedefinieerd; een analyse en een voorstel
B.J.G. van der Kooij
EUT/BDK/32 A conceptual Framework for Software Cost Control and Estimation
F.J. Heemstra, R.J. Kusters