On a tool for analysing cognitive processes using exploratory think-aloud experiments

Hans Heerkens*

Department of Operational Methods and Systems Theory,
School for Business, Public Administration and Technology,
University of Twente, P.O. Box 217,
7500 AE Enschede, The Netherlands
E-mail: j.m.g.heerkens@utwente.nl
*Corresponding author

Beatrice van der Heijden

Maastricht School of Management,
Open University of the Netherlands,
University of Twente, The Netherlands
E-mail: Heijden@msm.nl

Abstract: We develop a method that enables cognitive processes to be analysed quantitatively without having a conceptual framework in place at the start of the research. With current methods, the general structure of the cognitive process in question has to be known before quantitative analyses can be preformed. With the method, we studied the cognitive process of weighting the importance of attributes in a purchasing decision.

Our approach consists of seven steps:

- determining the research method (in our case: the think-aloud method)
- designing an experiment
- designing a data collection method
- designing a tool for preliminary analysis
- designing a tool for qualitative analysis
- designing a tool for quantitative analysis
- assessing external validity.

In this contribution we also provide a checklist for putting our method into practice.

Keywords: methodology of exploratory research; research on cognitive processes; methodology of think-aloud research; grounded theory; qualitative research methodology; attribute weighting; importance-assessment; decision making.

Reference to this paper should be made as follows: Heerkens, H. and van der Heijden, B. (2005) 'On a tool for analysing cognitive processes using exploratory think-aloud experiments', *Int. J. Human Resources Development and Management*, Vol. 5, No. 3, pp.240–283.

Biographical notes: Hans Heerkens is an Assistant Professor at the Department of Business, Public Administration and Technology at the University of Twente, The Netherlands. He teaches on Research and Problem-solving Methodology for undergraduate and PhD-students and Management in the Aerospace Industry. Hans also contributes to aerospace trade journals.

Beatrice van der Heijden is working as a Professor of Strategic HRM at the Maastricht School of Management, the Open University of the Netherlands, and the University of Twente. She graduated in 1990 from the Department of Work and Organisational Psychology at the Catholic University of Nijmegen and has a PhD (1998) in Career Development from the University of Twente. Her main research areas are: Career Development, Employability, and ageing at work. Currently she coordinates, among others, a cross-cultural European study on career success among ICT-professionals. She serves, among others, on the editorial board of SAM, *Advanced Management Journal, and Career Development International.*

1 Introduction

In psychology, the think-aloud protocol is an important method for information gathering on cognitive activities (Ericsson and Simon, 1993). This method can be used for both modelling cognitive processes and for testing hypotheses (van Someren et al., 1994). Yet, making sense of tens or hundreds of pages of think-aloud protocols is extremely difficult when an Apriori theory or a certain research perspective that can be used as a basis for structuring the protocols is lacking.

However, from a practical point of view, the think-aloud methodology can be very useful for theory building, since cognitive processes can be studied without having to specify the variables concerned in advance. The power of the think-aloud approach is that the research question can be rather open. Only a situation, in which the required cognitive processes can take place needs to be established, based on general ideas of the phenomena to be studied. Of course, the question of how to make sense of the rich but unstructured pool of data is far from trivial and this is an important issue that we want to address in this paper. The type of tool that one uses has an immediate impact on the sort of results that one can obtain. Often, purely qualitative descriptions do not provide answers that are precise enough to provide a sensible contribution to the scientific body of knowledge. Quantitative analysis on the other hand is often impossible as an analytical framework is lacking. Still, in this paper we want to show that useful compromises are possible, if an adequate preliminary analysis tool is part of the research methodology.

An example may illustrate this. In a study we aimed to find out how people come to a judgement on the importance of certain aspects of goods which they want to buy. From decision theory, we know elements that may be included in the judgement process. It is possible to express these elements in a mathematical formula for utility as a basis for rational behaviour. We also know which deviations from rational behaviour people are likely to make and which conditions influence their judgement, e.g., the number of items to choose from, or the amount of information that is available. But we do not know how people handle the elements that are included in the judgement process, i.e., the structure of the thinking process. We may know a little bit about the way these elements may

change during the thinking process, but knowledge on the kind of thoughts that induce these changes is lacking. In this sense this is explorative research.

It is possible to ask people about the way they make importance-judgements. We may interview them and ask them to memorise a decision which they made in the past. But if we do not know what we are looking for, what do we have to ask? We can ask open questions. But if people do not know how they think and why, what are they going to answer? They are likely to rationalise their thoughts in order to be able to communicate them to the researcher. Even experts in the field concerned may not be able to explain their way of thinking.

Thus, if we use methods like questionnaires or in-depth interviews, we may end up with results that are either very general, or not reliable, or both. The think-aloud method has been proven to be capable of revealing peoples' thoughts in full richness in a valid way (Ericsson and Simon, 1993) and to enable precise measuring. However, for precise results, as stated, we have to know in advance what we want to measure in terms of operationalised concepts. But in the early stages of exploration of a field, such operationalisations are simply are available. This makes the choice of a tool for preliminary analysis of think-aloud protocols rather tricky. Any tool is implicitly based on some model of reality. In order to be adequate for explorative research our tool should enable us to start from a *sufficiently general* model, incorporating notions on importance-assessment mentioned above. The emphasis on generality reduces the risk of overlooking interesting phenomena due to 'tunnel vision'. In a later stage, the descriptive power of the general model should allow us to obtain results that provide more precise insight into cognitive processes.

Hence, this paper focuses on the way in which think-aloud protocols can be analysed using only a very general structure beforehand while still yielding data that can be analysed in a quantitative way. The aim is to combine the richness of verbal protocols with the scientific rigour that is common in quantitative analysis. We want to apply this technique for problems for which no prior analytical framework exists. It is to be a generic scheme, adaptable to the research of a range of cognitive processes.

It must be stressed that our method is meant for studying individual actors only. The analysis phases within the method that we aim to develop may also be useful for research on group processes, but we did not evaluate its usefulness. And the measuring method (think-aloud sessions) is definitely not suited for research on groups.

So, the aim of this contribution is:

> To present a generic research design scheme while using the think-aloud method as the principal data gathering instrument, concerning problems pertaining to cognitive processes of individual actors that are only to a limited extent, embedded in an accepted body of knowledge, culminating in results suitable for qualitative and quantitative analysis.

Our research design scheme is based on our experience in a research project concerning a specific type of cognitive process: the weighting of attributes that describe alternatives in a decision process. Our starting point was not a methodological problem as such but the need to make a design for our research on weighting processes. We feel that the resulting generalised design as presented in this paper is applicable to research on cognitive processes in general under the conditions described in the above problem statement. Especially the role of a tool for the preliminary analysis of think-aloud protocols is not well recognised in our opinion.

The research design scheme is described in Section 2. We also discuss some main validity issues that have to be taken into account. In the next sections the research design scheme is illustrated for our research project on weighting of attributes. In Section 3, the think-aloud methodology that has been used in that study is described. Section 4 deals with designing a very loose framework for the analysis of the protocols, based on Simon's (1979) problem-solving model. Section 5 addresses the way the think-aloud protocols were analysed qualitatively, whereas the quantitative analysis is discussed in Section 6. The external validity of the results is addressed in Section 7. Finally, Section 8 concerns a discussion of our results and some recommendations for further research. In Appendices 1 and 2 we give some examples of results and the way we processed them. In Appendix 3, we give some 'tips and tricks' for researchers who want to use our method.

1.1 Scientific relevance

Our research design scheme is relevant for cognitive processes in general and management research in particular. Management research is seen as an inter-disciplinary science (Easterby-Smith et al., 1993; van Riemsdijk, 1999). Still, it is unclear how managers deal with interdisciplinary problems and how they mentally integrate chunks of incompatible information. We can measure input (information) and output (decisions), but the process that lies in between is a black box. Our method might be helpful in understanding the behaviour of individual managers engaged in decision-making.

The link between organisational phenomena and the cognitive processes of individual members of working organisations is not often studied, perhaps as the research methods for each part of the link – surveys, case-studies and simulations on the one hand and psychological tests, experiments, think-aloud sessions and measuring brain activity on the other hand – are often incompatible. In order to decrease this knowledge gap, the approach described in this paper is meant to allow the study of individual cognitive processes in organisational settings.

Our approach was derived from, and hence is specifically geared towards, cognitive processes in judging the importance of attributes when subjects have to choose between alternatives in a decision task. However, there does not seem to be any reason why the approach cannot be used to study cognitive processes pertaining to other tasks. That is to say, the research methodology that has been used seems suitable for cognitive processes in general.

2 A research design scheme for think-aloud experiments and aspects of 'design for validity'

We start with the various phases (in Roman script and capitals) that constitute the research design scheme. The main validity issues for each of the phases (in numbers) are described, in general. In subsequent sections, the phases and validity issues are dealt with more thoroughly and applied to our weight assessment research project. This section is concluded with a discussion of the new elements in our scheme, compared with literature.

Validity issues of the various phases may influence each other, creating an unjustified impression of overlap. We mention the most important relationships.

Phase 1: formulating the explorative research question and the choice for a think-aloud experiment

Cognitive processes are notoriously difficult to observe and measure, because the results of such processes are much more visible than the processes themselves. After all, a large part of any person's cognitive activities consists of automated subconscious thoughts. Verbalisation has its limits. This is reflected in the (internal) validity issues specific for this kind of research:

I. 1.1 verbalisation should lead to sufficiently rich and relevant information

I. 1.2 automated behaviour can be suppressed in a sufficient way.

Of course, in the data gathering phase, recorded verbalisation of the think-aloud experiment can be combined with other sources of information in view of specific requirements of the research question, see phase III.

Phase II: the design of the experiment

Three issues are in order here: the choice of the experimental setting, of the subjects and of the type of assignment.

As for the experimental setting in think-aloud sessions, the usual option is a laboratory experiment with individual subjects with which the cognitive processes can be studied in isolation from external influences like group interactions. This is in line with an overwhelming body of literature on psychological and decision research and it needs, in our view, no further elaboration. We summarise related validity issues as follows:

II. A.1 minimising process distortions by external factors, so that the information extracted (see I.1 and I.2) is actually about the process and not influenced by undue stimuli.

As for the choice of the subjects there is a direct relation with the assumptions underlying the research question and also, the external validity of results that one strives for. The following validity issues prevail:

II. B.1 qualification for the task to be solved in the assignment

II. B.2 the degree of experience with previous (analogous) tasks.

The qualifications usually refer to certain requirements on the expertise necessary for the task or inherent to a role given to the subject. In experiments where the research question is about routine tasks, the task in the assignment has a certain familiarity to the subjects. However, cognitive processes for non-routine tasks are even more interesting (yielding rich information, issue I.1) and challenging from a methodological point of view. Also, they pose less risk of automated behaviour (issue I.2). The familiarity of the subjects with the task should be closely controlled.

As for the structure of the assignment there is a strong relation of validity with avoidance of automated and/or enforced behaviour. Automated behaviour was covered in issue I.2. The amount of information and the way it is provided is crucial in this respect. Subjects performing non-routine tasks may need a certain amount of information, but care should be taken not to overwhelm them, especially not if the research is not about information processing. Note that this issue is not the same as II.B.1 and II.B.2: even well qualified and experienced subjects can be overwhelmed with information.

Furthermore, commitment to the task plays a role. Also, time pressure needs to be avoided, since this may hinder the subjects in verbalising their thoughts (Ericsson and Simon, 1993). Altogether this leads us to the following validity issues:

II. C.1 the artificiality of the task/role should not lead to a-typical behaviour or non-commitment

II. C.2 stress and time pressure control, in order to prevent induced behaviour and facilitate verbalisation

II. C.3 control of the supplied information.

Phase III: data collection design

In think-aloud experiments, audio recording of the sessions is usually the basic material. In addition, a method that answers to specific data requirements needed for the research question can be chosen.

Such additional protocols can take various shapes, like observation protocols, notes made by the subjects themselves, notes made by the experiment supervisor (e.g., if some form of interviewing is used), recording of computer actions or video recordings. Altogether there are several validity issues concerning the interaction of the subject with the recording medium or an interviewer, in addition to external effects as covered by II.A.1. These can be summarised as follows:

III.1 control of interference with the medium and/or experiment supervisor.

Phase IV: the design of a tool for preliminary analysis

The input of this phase is general knowledge about areas related to the research question, plus the as yet unstudied protocols. The more the research has an explorative character, the less the variables and relationships to be studied are defined. The output is, in general, a model of the cognitive processes to be studied, culminating in a coding scheme in which variables and concepts are available that can be used as the starting point for the qualitative analysis. The coding scheme will, at this stage, merely refer to identifying potentially relevant variables and concepts, not to establishing relationships between them or measuring quantities or variables. In this sense the analysis is preliminary to the following phases.

Especially in exploratory research, a basic framework, a model for analysis is needed to make sense of the think-aloud protocols. This model provides the elements to look for in the think-aloud protocols. In exploratory research sufficient generality of the model is a main criterion. If not, interesting phenomena may be ignored from the outset. Generality has repercussions in terms of lack of detail. The model is not necessarily complete and it does not necessarily relate the various elements to each other. Furthermore, the model should be capable of handling the dynamic perspective inherent for a process description, up to a certain degree. In case of a static perspective, various static elements (say, input and output elements of reasoning) can be captured but not, to any degree, the activities that lead to changes. Of course evolution of cognitions is a characteristic of cognitive processes in think-aloud experiments.

This framework should form the basis for further qualitative analysis and it should also preferably yield some basic quantitative information, like, say, how many subjects performed specific kinds of activities.

Validity issues in this respect can be summarised as follows:

IV.1: construct validity

This issue refers to picking the right elements from a body of knowledge concerning the research question and framing them in a sufficiently general model. This is a challenge if there is little or no previous research available to use as a guideline. Then, picking the right elements is an art; if the researcher does not have the talent no interesting or relevant picture in terms of results will emerge.

IV.2: defining the various elements in the general model

This should be done in such a way that they can be unambiguously identified in the protocols. If the problem given to the subjects is unstructured and the assignment does not prescribe a solution method, the subjects have a lot of freedom in the execution of the task. The more freedom for the subjects, the more difficult unambiguous identification of the model elements is likely to become. Therefore, it is non-trivial to define the elements of the model in such a way that they are general enough to discover regularities across the protocols, and yet so specific that they can give meaning to the text of the protocols.

IV.3: coder consistency

In line with the previous point, coder consistency is a challenge as well. Generality in the model may easily reflect itself in difficulties to define variables and indicators so clearly that a coder would give the same code to identical (but sometimes differently formulated) variables over time. The elements in the general model should be straightforward enough for operationalisation, to ensure correct interpretation by the coder. The analysis tool has to lead to identical labels to similar phenomena, even if they manifest themselves in different ways across and within protocols. This does not mean that two or more coders should code the same phrase identically; the classic meaning of coder consistency. In this phase, one person does the coding, so that he or she can develop experience and a 'feeling' for the protocols at hand. So, coder consistency here means that the coder remains consistent during the coding process. Naturally, as experience and insight in the protocols is gained, consistency may suffer. This has to be addressed by, for example, going through each protocol several times. In phase 6 coder consistency will be addressed again, then in the more common meaning of inter-coder reliability.

IV.4: cognitive modelling restrictions

The analysis tool should not restrict the interpretation of the protocols, in the sense that relevant information in the protocols would go unnoticed in the qualitative analysis. There is the danger that the researcher focuses too much on labelling identifiable elements in the analysis tool with clear, but too restrictive, meanings, thus leaving out some other interesting aspects. This can, of course, be at odds with IV.3, in the sense that clear pre-defined elements of the general model may reduce the urge to look for non-trivial notions in the protocols.

Phase V: the qualitative analysis

In exploratory research this can be done according to the grounded theory approach (Strauss and Corbin, 1998). This is a stepwise method for building rather than testing theories. The steps are described in Section 5.2.

Basically, this method involves studying the think-aloud protocols a number of times. One of the aims is to refine the preliminary analysis tool until sufficient confidence in its

validity issues is built up. Next, the essence of science is addressed to identify relevant phenomena that have not been identified in the general model of the previous phase, and to establish some (qualitative, logical) relationships between those phenomena. Again, but with more structure than before, the resulting model should be able to deal with a dynamic perspective of the activities which the subjects performed during the cognitive processes (e.g., their sequencing) and the relationships instead of describing merely static behaviour. Typically, the dynamic perspective can reflect itself in recognising phases in the cognitive process under study consisting of clusters of task related activities and establishing relationships between these phases.

Validity issues are partly the same as for the previous phase (designing the preliminary analysis tool), but in addition:

V.1: objectivity in establishing relationships between phenomena

In the Grounded Theory approach, there are no fixed criteria to establish whether relationships exist. The method is meant for exploring relationships in a qualitative way. The relationships found need to be tested in quantitative research. Note that this goes beyond coder consistency (IV.3) because it is not about coding but about the analysis of the coded material.

The input of this phase is the content of the protocols from the experiment (see phases II and III) after applying the preliminary general coding scheme (see phase IV). The output is a qualitative model of the cognitive process to be studied from a dynamic perspective.

Phase VI: the quantitative analysis

The aim of the quantitative analysis is to quantify some of the results of the qualitative analysis. For example, if the qualitative analysis leads to recognition of phases consisting of clusters of activities, then in the quantitative analysis, we may want to find out how much effort the subjects devote to each cluster. Besides, we may want to check the main results of the qualitative analysis in terms of significance. The input of this phase is the qualitative model of phase V, on the basis of which a formal coding scheme for measurements in a dynamic perspective is based. The output consists of results on categories of phenomena, structured and stored in such a way that statistical operations on them are possible.

Whether it is possible to establish statistical relationships depends on the number of subjects in the think-aloud experiment. But even if the number of respondents is (far) too small for statistically sound conclusions, quantitative analysis is worthwhile, since it provides more precise information on what to expect in a generalised context. In this paper we shall not elaborate on the statistical validity issues. They are well covered in a wealth of literature on empirical research (see for an overview Cooper and Schindler, 2003). Additional validity issues are:

VI.1: content validity of the named categories of phenomena

This concerns the relationship between what we measure and what we propose to measure given the name and inherent interpretation of the category. This is an issue for, despite the work done during the design of the general model (issue IV.2) and the qualitative analysis, further operationalisation of coding categories was sometimes necessary and protocols were so diverse that the relationship between phrases in the protocols and the categories to which they seemed to pertain, were not always clear.

VI.2: inter-coder reliability

The issue is to what extent did different coders give the same codes to a category of phenomena. It is not the same issue as coder consistency (IV.3) when designing the general model.

- the data is coded anew during the quantitative analysis and the issue has to be revisited
- when designing the general model there is only one coder (who has to be internally consistent), whereas in the quantitative analysis phase there are at least two (who have to be consistent with each other).

It is always an issue with coding, but in the case of an unstructured nature of the assignment the subjects have to fulfil, it calls for extra attention.

Phase VII: assessment of the external validity of the final results

Up to now, the focus was on the internal validity issues in the think-aloud experiment. Of course there is a claim for a more general scope of validity of the final results than the experiment itself. This gives rise to the following validity issue:

VII.1: well-founded external validity

Of course, this issue should be anticipated already in the earlier phases of the research design discussed above, although the emphasis was on internal, construct and content validity; a prerequisite for external validity. But, as a final step after execution of the research design, reflection on external validity is necessary. It is not uncommon, that this leads to further interesting research questions.

What is new about our research method

Our method is innovative in three respects.

- It provides a structured way to obtain exact, even quantitative information about cognitive processes from think-aloud protocols in case of explorative research without a specific analytical framework being available beforehand. There are numerous generic methods available (see, for example, Chi et al., 1988; Eisenhardt, 1989; Strauss and Corbin, 1998; van Someren et al., 1994). However, they either take certain general models as inputs (van Someren et al., 1994), are more directed towards analysis of interview protocols (Strauss and Corbin, 1998), or they are not directed specifically at analysing cognitive processes but merely in measuring their results (Chi et al., 1988; Eisenhardt, 1989).
- It provides a way for analysing protocols from experiments where subjects have complicated tasks for which no pre-defined solution procedures exist. In real life, we often encounter such non-routine tasks. Accordingly, a method to analyse them in-depth is valuable. Many think-aloud experiments concern more or less structured tasks (see for an overview Ericsson and Simon, 1993).
- The third innovative aspect is the emphasis on a general model as a means for preliminary analysis of the protocols in the think-aloud experiment. This general model is not our definitive analysis model, but it provides a general idea of elements to look for in the protocols, a sort of 'checklist'. The use of such a list of general elements to look for proved to be a valuable asset compared with starting from a completely blank sheet (if this were possible at all), if an elaborated model is not available.

Our method is more a combination of existing methods (with an emphasis on the interface between the think-aloud method and the grounded theory, the tool for preliminary analysis of the protocols), rather than a completely new method. Still, it opens up areas for which, as far as we know, a sound and practical research method was not available until now.

This concludes the general description of our research design. In the following sections, the way the research design scheme can be used in our research on attribute weighting, how that research project was conducted and the way the validity issues were handled are discussed more thoroughly.

3 The think-aloud method in attribute weighting

3.1 Pros and cons of the think-aloud method

Let us now illustrate the logic behind the research design scheme in more depth with our research project on attribute weighting. The (summarised) problem statement of the research can be formulated as follows:

> How do individual actors within an organisational context arrive at importance-judgements when involved in the initial phase of buying a capital goods, in cases when this is a non-routine decision for the actors?

The first phase of the research design scheme asks for reflection on whether a think-aloud experiment fits well with this research question (research design scheme, phase I). Section 3.1 starts with an outline of the reasons for choosing the think-aloud method as the general format for our study. In Section 3.2, the way in which the think-aloud experiment was designed is discussed with regard to the assignment, the choice of the experimental setting and the choice of the subjects (research design scheme, phase II). Also, some details on our data collection are given (research design scheme, phase III) in Section 3.3.

3.1.1 Phase 1 of our method: the choice for the think-aloud method

The validity issues addressed in Section 2 can be concretised to fit our particular problem statement. The potential to elicit rich information means (I.1), in our case (see the problem statement above), the ability to extract information about cognitive processes from individual actors in an organisational context (which may be simulated). The need to suppress automated behaviour (I.2) implies that it is not only the results of cognitive processes that need to be measured, but the processes themselves. The danger of automated behaviour was low, since we focused on non-routine decisions. But we also wanted the subjects to have so little experience with the task that they could not revert to automated behaviour. All in all, this led to the following requirements.

The research method should be capable of:

Requirement 1: *Being used to study individual actors. It does not have to be suitable for studying groups.*

Requirement 2: *Being used in either a 'real-life' or a simulated organisational context. It does not have to be suitable for studying subjects in a private environment.*

Requirement 3: *Being used to study evolution during cognitive processes, not just the verbal or behavioural results of these processes. This, of course, leaves open the possibility that verbal or behavioural phenomena serve as indicators for the thought processes.*

Requirement 4: *Being used in a situation where the thinking and behaviour of the subjects is not regulated by formal procedures or constraints, as may be the case in, especially, the preliminary phases of a non-routine decision process. Subjects can use any concept of 'importance' they desire and perform the importance-assessment in any way they find appropriate.*

Requirement 5: *Being used with subjects who have little or no experience in the task at hand. Our research concerns non-routine decisions, in which no previous importance-judgements are readily available. This implies that making a new judgement is both necessary and relatively challenging. It is therefore doubtful that the subjects will use standard methods for making importance-assessments, even if such methods were available. The consequence is the same as in the previous point: the research method has to be able to capture a potentially wide array of variables and structures of assessment processes.*

As will be explained shortly, the think-aloud method fulfils all the above requirements. In its simplest form, the think aloud-method is nothing more than letting an individual subject verbalise every thought at the time and in the form in which it occurs to him or her (Davison et al., 1997; Ericsson and Simon, 1993; Séguinot, 1996; van Someren et al., 1994). The verbalisations are recorded on tape and typed out partly or completely for analysis. The typed-out version of the verbalisations is called the think-aloud protocol. Cognitive processes take place in a person's mind and hence cannot be observed directly. Certain behaviour, for example the choice of a particular goods from several alternatives or a stated importance-judgement, can be seen as results of cognitive processes. But these processes themselves remain largely invisible, i.e., a 'black box' in terms of Séguinot (1996). Since our research aims to analyse the reasoning that takes place within the 'black box', it is logical to choose the think-aloud method as a means of gathering data.

Let us now consider the issue of verbalisation in the think-aloud method (I.1). In comparison to another, very popular, verbal method, retrospective reporting, the think-aloud method offers several advantages.

- that the subjects do not have to have insight into their own mental processes, since they do not need to explain their thoughts
- subjects can be given a task to perform and asked to think aloud while doing it; so there is no need for prior experience with the task, because no reporting on previous experiences is sought

- it is difficult for the subjects (although perhaps not impossible) to rationalise their thought processes, i.e., to make them look more rational than they actually are
- the think-aloud method enables detailed analyses of how subjects behave through a problem-solving process (Ericsson and Simon, 1993); in this case an importance-assessment process.

Yet, the think-aloud methodology does not come without a price. It has occasionally been used to analyse complex cognitive processes (Davison et al., 1996), but Ericsson and Simon (1993) give many examples of think-aloud studies concerning the solving of highly structured problems, like mathematical problems or variants of the tower of hanoi problem. These problems are characterised by the fact that the number of possible actions is limited (for example, in mathematical problems there may be actions like adding, subtracting and the like), and that the solution can be judged to be right or wrong. In our research, there are an infinite number of possible actions and there is no 'right' solution (see Section 4). This makes both the coding and the interpretation of the data difficult.

Processing and analysing data from think-aloud studies is extremely labour-intensive. This is addressed in the literature (Whitney and Budd, 1996) and definitely proved to be the case in this research. The number of subjects that can be handled is therefore limited if an in-depth analysis is required.

Another drawback of the think-aloud method is that it is applicable only in certain situations. It is not usable for analysing group processes, since thinking aloud and communicating with others do not go together very well. This is one of the reasons why this research focuses on individual importance-assessment processes, the other being that we were unable to isolate determinants for individual reasoning from the influence of group processes. Also, the think-aloud method can only be used for analysing processes at the moment they occur. There are other pros and cons of the think-aloud method that have been discussed extensively elsewhere (Ericsson and Smith, 1991), but these have no consequences that are specific for this research, and therefore will not be addressed.

When the characteristics of the think-aloud method are compared to the requirements mentioned before, the following observations can be made. The method focuses on individuals (Requirement 1). The assignment that subjects have to fulfil can either take place within an organisation (for example, letting a manager think aloud about a decision to be made) or the organisational context can be incorporated in the assignment (Requirement 2). Although the thoughts expressed by a subject are only those thoughts that are verbalised, the general structure of the thought process is likely to remain intact as long as the method is used properly (Ericsson and Simon, 1993), and there is no better method available to study thought processes (Requirement 3). The assignment can be either restrictive or open, as long as it is clear enough for the subjects to organise their thoughts and verbalise them procedures and other constraints can be controlled by the design of the assignment (Requirement 4) and the method can be used with inexperienced subjects (Requirement 5).

The assessment is that even though the think-aloud method has some drawbacks, these can be overcome. The detailed insight into cognitive processes that we strive for in the research project on attribute weighting and the possibility to give tasks to subjects under controlled conditions far outweigh the drawbacks, which in any case, could be neutralised to a significant extent.

If we now look at the validity issues pertaining to this phase of our method.

3.2 The think-aloud experiment as applied in our study

This section pertains to two of the phases 2 and 3 of our method: the design of the experiment and of the data collection method. It is structured according to the validity issues discussed in Section 2.

3.2.1 The setting (phase II.A)

In our experiment we chose a laboratory setting at our university. Even though an organisational context is part of the assignment, the fact that we worked with students was decisive in this respect.

3.2.2 The subjects (phase II.B)

The 18 subjects in our study were seventeen third and fourth year undergraduate students and one MBA-student from the University of Twente, Enschede, the Netherlands. All but one of the undergraduate students had studied management science at the School of Business, Public Administration and Technology. The MBA student had a technical MSc. It is a natural assumption that this kind of student has affinity with and commitment to the assignment as described (II.C.1) and it is clear that these subjects should be qualified for the task in the assignment (II.B.1).

Based on studies on alumni, one can assume that some subjects are going to attain positions wherein acquisition decisions are to be made, over the course of their future career. They were also assumed to be sufficiently capable of emulating the view of a person within an organisation having to perform an importance-assessment, since they had conducted several practical case projects during their study. Interviews that were conducted after completion of the assignment indicated that this assumption was by and large correct. By using students and not acquisition experts, we avoided 'automatic' thought processes (Ericsson and Hastie, 1994), that do not enter working memory and hence are not verbalised. Our second motive for choosing students and not experts is expressed by Harris (1992), cited by Séguinot (1996), in the context of think-aloud studies: "To study advanced forms of a skill before understanding how beginners do it is to build the house before digging the foundations". Hence in this sense experience was clearly controlled (II.C.2).

3.2.3 The assignment (phase II.C)

During the think-aloud sessions, each subject had to execute an assignment. The subjects had to imagine that they worked for a travel company that transported passengers to Amsterdam's Schiphol Airport, some 100 miles away from the city where the students' university is located (University of Twente, Enschede). The students were asked to advise the management regarding the acquisition of a new fleet of minibuses to replace the current one. The advice they had to give was an importance-judgement of safety vs. passenger comfort. It was stressed that the subjects were allowed to perform the assessment in any way they liked and that there were no limits as to the kind of reasoning that was allowed. Since all subjects, but one, had a background in business studies we can assume that artificiality of this sort of assignment is not an issue for them (II.C.1).

The only condition that was communicated to them implied that they would have to be able to explain their judgement to the management of the company afterwards as if this were a real-life task. This did not mean that the management had to agree with it. It was also stressed that there was no choice between types of minibuses to be made. The subjects had 1.5 hours to complete the task, which was more than sufficient for everyone. Hence time pressure was not present (II.C.2).

A problem that came up was that, in order to understand the assignment, the subjects had to absorb a greater quantity of information than we would have liked. Our solution was to present the information in pieces. First, the subjects had to read a short text on the general purpose of the assignment. Then, some information on the company and the decision context followed. By now, the subjects knew that they had to make an importance-assessment concerning minibuses, but they did not yet know about which attributes. Also, they were given two brochures about minibuses, so that they had a better idea of what a minibus actually was. They were given 15 minutes to read through and process the information. Finally, they were given the attributes about which they had to make an importance-assessment. Hence, information control was a clear issue in this assignment (II.C.3).

3.3 Data collection (phase III)

In our research project, we used audio recording of the sessions. Interaction with the experiment supervisor was kept at a minimum to avoid interference (III.1). In order to get acquainted with thinking aloud, the subjects performed three exercises before commencing the assignment, as suggested by Ericsson and Simon (1993).

Afterwards an exit interview was conducted with each, subject wherein issues like their confidence in the quality of their work were addressed.

4 A general model for preliminary protocol analysis concerning attribute weighting processes (phase IV)

The development of our analytical framework for analysing attribute-weighting processes started with Simon's general problem solving model, which will be discussed in Section 4.1. This leads to a general model of the importance-assessment process (Section 4.2), operationalised in a preliminary coding scheme (Section 4.3). The validity of this approach is addressed in Section 4.4.

4.1 The starting point: Simon's general problem-solving model

Having to give an importance-judgement can be seen as a problem, the way to its solution being the importance-assessment process. There are hundreds of problem-solving models, many of them prescriptive, some descriptive (see for an overview Hicks, 1995). The problem-solving model of Simon (1979) is very general and directed at information processing by individual problem-solvers, thus fitting in the context of our problem statement.

Simon describes problem solving as 'moving through a problem space' (Hunt, 1994; Newell and Simon, 1972; Simon, 1979). It is clear that this view accommodates a dynamic perspective from the outset. The problem space is the way in which the

problem-owner (the person given the task of solving the problem) sees the task environment (the task at hand). The problem-owner is seen as an information-processing system, going from one node in the problem space to another. In the case of a complex, unstructured problem, the problem space may change as the problem-solving process evolves. A problem space contains all kinds of solutions to a problem and all the results of steps made during the problem-solving process, insofar as they are results of 'legal moves'. A legal move is a transition from one step in the problem-solving process to the next that is in accordance with the problem definition. It might be that not all elements of the problem space are relevant for the solution of the problem, but they are all permitted as results of legal moves. Of course, legal moves capture the essence of the process dynamics, as well as the essence of 'solving'.

In our study, the importance-assessment process is seen as a movement of an actor through his or her problem space. Therefore, it is essential to develop a way of representing the problem space of these actors.

4.2 A preliminary model of the importance-assessment process

4.2.1 The problem space

Of course the problem space depends on the sort of problem considered: in our case it refers to elements associated with importance-assessment processes. Essential elements will be attributes and weights. Their background in scientific theory will be discussed shortly. As for cognitive processes arguments are introduced as another element in the problem space. In the dynamic perspective of legal moves made during the problem solving process in importance-assessment, we can now identify:

- The attributes that the actor starts with.
- All the modified attributes that could possibly be developed from the original attributes.
- All the weights and weight ranges that could possibly be assigned to any attribute of the problem space.
- All arguments that could possibly be used for justifying all possible attribute-weight combinations.
- All possible attribute-weight combinations, plus their associated arguments. Note that not all arguments that could logically be associated with these attribute-weight combinations have to be actually associated with them. This is up to the actor. We are not concerned with whether the assessment process or importance-judgement is 'logically' correct. The importance-judgement is a subset of these attribute-weight-argument combinations.
- All possible forms of the utility function used.

The above-mentioned elements form the basis for discovering regularities in the evolution of cognitive states that actors go through when performing an importance-assessment. They do not yet provide a list of cognitive operations but merely indications of where to find them. In Section 4.2.2, an initial classification of legal moves is given, which was further developed during the qualitative analysis of the protocols. Since the importance-assessment process is highly personal, we have no outside norms

for legality of a move. Hence in this sort of research on cognitive processes, if a person makes a move, we accept it as a legal move. Hence, all moves are legal. Still, we retain the term 'legal moves', in order to follow Simon's terminology.

To derive the basis for the problem space of the importance-assessment process, we refer to the elements as recognised in the theoretical concept of a utility function. In this concept there is a set of attributes on which an alternative that a decision maker considers, is scored. The scores on attributes are aggregated into an overall utility of an alternative. The relative importance of each of those attributes is called its weight. (Keeney and Raiffa, 1976), and the simplest utility function is the linear additive function, which can be expressed as:

$$U_i = \sum_{j=1}^{N} A_{ij} W_j$$

where U_i is the utility of alternative i, A_{ij} is the score of alternative i on attribute j and W_j is the weight of attribute J (identical for all alternatives).

Concretely, the score on each attribute of an alternative is multiplied by the weight of that attribute and the results, called 'partial utilities', are added to get the total utility, or attractiveness, of an alternative. The alternative with the highest attractiveness should logically be chosen. An example as used in our study is given in Table 1. A decision-maker has to choose which of two types of minibuses he should purchase for his travelling company.

Table 1 The utility of two types of minibuses

	Score on safety (weight = 0.4)	*Score on comfort* (weight = 0.3)	*Score on running costs* (weight = 0.2)	*Score on price* (weight = 0.1)	*Total utility*
Type A	3	5	1	2	3.1
Type B	4	4	2	5	3.7

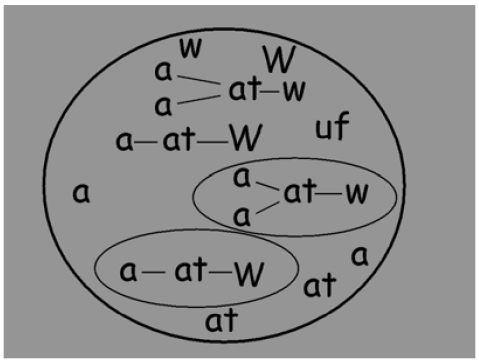
In this case, the type B bus should be chosen for it has the highest total utility.

When performing an importance-assessment, it is obvious that arguments are given for the various cognitive operations that are performed, such as attributes that are considered to be relevant or weights that are assigned. Note that, as will be the case in the remainder of this contribution, 'arguments' here do not only stand for single arguments but also for chains of interrelated arguments.

Weights and scores need not be numerical values, but may be expressed as fuzzy ranges or in a qualitative way in the problem space. For example, a weight may be 0.3, but may also be somewhere between 0.3 and 0.4, or may be given in a qualitative way like 'very important' or 'not important'. So, their measurement level can vary from nominal to ratio, (Blalock, 1981; Swanborn, 1973; Swanborn, 1987). Our framework will accommodate this generality.

Figure 1 shows our problem space for an importance-assessment process and its associated weight judgements. In this figure we can find all elements listed at the start of this section. The encircled weight-attribute-argument combinations represent the combinations eventually included in the importance-judgement. The other elements are used at some stage in the decision making process but are not included in the resulting judgement.

Figure 1 The importance-assessment process



uf: shape of the utility function.

at: attribute.

w: weight value.

W: weight range.

a: argument.

It is clear that this problem space is infinitely large. There are an infinite number of weights, even if weights are set between 0 and 1. Also, the number of arguments in favour of or against incorporating a certain attribute and any weight for any attribute that an actor may consider is infinite. No actor can oversee the entire problem space, but no actor needs to. The theoretical problem space, in contrast to the actual problem space, is not what the actor has in his or her mind, but rather the elements that the actor could pick in the quest for an importance-judgement.

How does the above help us in constructing a general model for preliminary analysis of the think-aloud protocols of importance-assessment processes?

- it shows possible elements that mental activities of the actors (legal moves) may be related to. This is further discussed in Section 4.2.2
- the different combinations of elements may point to phases in the assessment process that can be distinguished, for example: the definition (and elimination) of attributes vs. their weighting.

Now that the problem space of the importance-assessment process has been defined, it is time to look at some legal moves of this process. We will not discuss all the legal moves that we derived from the model, but confine ourselves to some examples that give an insight into the way our research method works.

4.2.2 A classification of legal moves

The general categories of the legal moves we can apriori identify, are in line with Simon two sub-processes for the solution of complex problems, i.e., "an understanding process that generates a problem space from the text of the problem, and a solving process that explores the problem space to try to solve the problem" (Simon, 1979, p.268). The first sub-process is often called 'structuring the problem'.

4.2.2.1 Structuring the problem

Simon (1979) indicated that, whereas structured problems like the tower of hanoi problem are clearly defined in terms of the starting situation, legal moves and solutions (and hence problem space), ill-structured (wicked) problems may need structuring before the search for a solution can start. The structuring process can pertain to:

- the starting situation, like further definition of the attributes to be weighted
- the result, for example the sort of importance-judgement desired (qualitative vs. quantitative, the extent of exactness), which sort of arguments are likely to be considered sufficient or legitimate
- the available means, like information and the organisational context
- the legal moves to be employed.

The main legal moves of this phase; choosing the shape of the utility function and processing attributes, are discussed below. Other legal moves are not addressed because they are not unique to the importance-assessment process.

Choosing a general form of utility function in the problem space

In no research that we are aware of, have subjects been asked to explicitly choose a general type of utility function (for example, an additive or multiplicative, linear or non-linear function) before, during, or after the weight assessment process. But it is possible that an actor, familiar with the phenomenon of utility functions, explicitly or implicitly chooses such a function before starting the weighing process. This possibility will be taken into consideration.

Processing attributes

The notion of the processing of attributes comes from two sources: the obvious necessity of comparing attributes in order to determine their relative importance, that is to say, to express them in some sort of common denominator and the notion of cognitive processes as a series of steps in which one step forms the input for the next one. A logical way to analyse the processing of attributes would be to divide them into a series of steps in which attributes are progressively modified until they reach the stage in which comparison is allowed. In Simon's model, the output of one legal move (a modification of an element of the problem space), in this case an attribute, forms the input for the next legal move. The notion of describing processes as a series of transformations where the output of one transformation is input for the next is also well known in systems theory (de Leeuw, 1997).

Ways of processing attributes

The following ways of processing attributes have been identified, both from attribute characteristics described in, for example, methodological and statistical literature (for example, Blalock, 1981; Cooper and Schindler, 2003; Swanborn, 1973; Verschuren, 1980), and from research on attribute judgements. We restrict ourselves to ways of processing that we expect to be reliably identifiable in the protocols.

- *Splitting attributes (see, for example, Borscherding et al., 1995; Póyhónen and Hámálánen, 1998)*. 'Safety', for example, can be split up in sub-attributes that make a minibus safe, like 'strength of the chassis' and 'the availability of seatbelts for passengers on rear seats'. It can also be split up in several effects on passengers like 'the number of accidents' and 'the number of fatalities per accident'.
- *Integrating*. This is the opposite of splitting. An actor may take several sub-attributes together in one attribute.
- *Concretisation of attributes*. This means lowering the level of abstraction. For example, safety may be defined as 'the chance that a minibus arrives at its destination without being involved in an accident'. The difference with splitting up is that the whole attribute is thought to be covered by the newly formulated attribute.
- Abstraction as the opposite of concretisation.
- *Re-formulation*. This implies giving the attribute a new name without effectively changing the measurement unit, or it implies changing the measurement unit without affecting the measurement level, the relationships with sub-attributes (1 and 2) or the level of abstraction (3 or 4). For example, an actor may rephrase "availability of a stereo set" as "does the minibus have a radio?" or "is the braking distance (assumed to be in meters) better or worse than average" as "how long is the braking distance"?

Changes in the way subjects describe attributes and their use during the think-aloud sessions will be used as indicators for the ways in which those attributes are processed.

4.2.2.2 Solving the problem

Absolute vs. relative weighing (pertaining to the weights and attributes in the problem space)

Absolute weighing means that weights are given to each attribute in isolation, without comparing the importance of the various attributes to each other. Relative weighing means that the subject relates the weights of attributes to each other. The distinction between absolute and relative weighing can also be found in the various types of elicitation methods commonly used (see, for example, Harte and Koele, 1995; Jaccard et al., 1986; Saaty, 1980).

Timmermans (1993) uses the distinction in absolute and relative scoring of attributes. An example of absolute scoring is 'this minibus has a high level of safety'. Its relative equivalent would be: 'The Opel seems to be safer compared with the Volkswagen'.

Holistic vs. dimensional weighting (pertaining to the weights and attributes in the problem space)

Timmermans (1993) distinguishes between holistic and dimensional judgement. A holistic judgement covers the alternative as a whole; i.e., "this is a very attractive minibus". A dimensional judgement covers only some attributes, i.e., "this minibus has comfortable seats". Likewise, in this paper, a distinction is made between holistic weighting (making a holistic importance-judgement), where an attribute as a whole, gets a weight, and dimensional weighting, where any or all sub-attributes of an attribute get weights. In the experiment performed in this research, subjects were supposed to perform a holistic importance-judgement, but as splitting up attributes is a way of problem structuring (see above), dimensional importance-judgements are possible.

Linking arguments to attribute-weight combinations (pertaining to attributes, weight and arguments in the problem space)

Often, this legal move will be performed in combination with one or more of the previous ones. An actor may link an attribute to a weight value, or a weight range, and then provide arguments for this action. As with the preceding legal moves, the result may be provisional or final. The actor may consider (groups of) arguments in favour or against the importance of an attribute without linking them to a specific weight value.

Many types of arguments for weights can be identified. One type of argument that stands out in the literature, handling risk (Kahnemann and Tversky, 2000; Keeney, 1992). Because it was impossible to overlook and hence was one of the perspectives from which the initial analysis of the protocols took place, we included this type of argument in our classification of legal moves.

Now that the main elements of the importance-assessment process have been identified, a preliminary coding scheme can be developed as the starting-point for the analysis. In the next section, the design of the coding scheme is described.

4.3 The tool for preliminary protocol analysis

The primary aim of this tool is to serve as a preliminary coding scheme. It has to identify some of the elements that have been discussed in the previous sections, so as to get some structure in the 'pile of data' that the protocols represented. The coding scheme was further developed during the qualitative analysis, and hence provided a way to structure the results of this analysis. The word 'preliminary' needs some explanation. It is used because the coding scheme developed for and during the qualitative analysis was, for the most part, rather general and was used more to identify and structure phenomena than to provide precise or so-called 'hard' results. It was used more as a structuring tool than as a measuring instrument. It was 'preliminary' relative to the coding scheme used for the quantitative analysis.

The first part of the coding scheme dealt with identifying the (sub) attributes used by the subjects and the extent of (sub) attribute processing. We started with underlining every (sub) attribute used by a subject and then establishing whether each (sub) attribute was the result of decomposition, integration, specification, abstraction or re-formulation of a previously mentioned attribute. As the subjects seldom identified processing activities explicitly, they had to be inferred. So, a set of rules was developed to separate, for example, specification from decomposition.

An example of such a rule is: if attribute X1 is processed into only one new attribute X2 at a lower level of abstraction and it is clear that the subject sees X2 as covering X1 entirely, it is a case of concretisation. If it is clear that the subject feels that X2 only partly covers X1 and hence that more sub-attributes are needed to completely cover X1 (regardless of whether the subject actually mentions these other sub-attributes) it is a case of decomposition. In this way, coding rules were developed for all processing activities. Appendix 1 shows the result of the complete coding scheme for one particular subject. Although this 'attribute-processing scheme' provides some quantitative data, such as the number of processing activities, it was meant to be largely descriptive and served as a basis for the qualitative analysis only.

With this scheme at hand, more information could be extracted from the protocols. For example, the instances when weights were given to attributes could be made explicit. This made it possible to recognise not only absolute, relative, partial and holistic weighing, but also two kinds of weighting that would eventually emerge as phases in the model, i.e., homogeneous and heterogeneous weighting. The former means weighing two sub-attributes of the same main attribute against each other. The latter means: weighting sub-attributes of different main attributes. Examples are, respectively, the weighing of head-room and leg-room (both comfort) and the weighing of the quality of the braking system (safety) and the amount of leg-room (comfort).

Looking at the weights that were given might give an idea of the shape of the utility function, if any, that the subjects used. For example, all subjects used weights that were independent of the score of the attributes. As there were no indications that scaling techniques were used, it seems fair to conclude that the subjects either used no utility function at all, or used a linear additive function.

In sum, with the attribute-processing scheme, three elements of our model (attributes, weights and the shape of the utility function) have been addressed, plus the attribute processing activities and the various types of weighing. 'Addressing' does not mean 'analysing in-depth'. Besides, not all elements of the model have been addressed yet. Rather, depicting the attribute scheme for each subject is only a preliminary step to the start of the further qualitative analysis.

In the next section, the qualitative analysis, in which the second step of the development of the coding scheme took place, is described. But first we apply the validity issues identified in Section 2 to the development of our general mode as described above.

4.4 Validity issues concerning the general model and the coding scheme

The validity issues numbered IV.1–4 are relevant here. By taking Simon's general problem-solving model and combining it with literature on decision theory, we hope to have maximised construct validity (IV.1). The construct validity in this research probably was not lower compared with other exploratory studies. The lack of relevant literature was one of the reasons for using the grounded theory approach in the next part of the analysis (see the next section).

Defining variables or phenomena unambiguously (issue IV.2) proved difficult. In the end, by crosschecking the protocols (did a certain criterion for distinction that was suitable for protocol one also hold for protocol two), we believe that we have achieved a satisfactorily result. But in case other groups of subjects had been studied, for example, experts, the rules may have to be adjusted.

We took further measures to handle ambiguity.

- In the coding scheme, we included the segments from the protocol that pointed at a certain phenomenon, and checked the coding after some time to see whether our insights had changed. Also, comparing citations from different subjects became relatively easy this way. This frequently led to adjustments in the coding.

- We did not try to infer what a subject might have meant to say, but only what he or she actually said. This seems logical, but sometimes the temptation to interpret before coding was strong. Refraining from inferring may, of course, have led to either incorrect coding or dismissal as irrelevant of certain phenomena. The latter was countered by the fact that there was a lot of repetition in the protocols, a variable that was missed on one occasion was almost sure to be spotted on another, even if was formulated slightly differently.

On the whole, we feel that ambiguity was sufficiently low for the further qualitative analysis. Our procedures for safeguarding against too much ambiguity also, were checks on consistency (IV.2). This holds especially for cross-checking the protocols and re-studying them over time. As the style of formulation of the subjects was so varied, our iterative way of working was bound to highlight at least some of the inconsistencies in the coding.

Coder-independence (issue IV.3) was a thorny issue, for the coding required much experience. Solutions that were implemented were to have the more simple coding activities checked by an assistant. Another solution was to check for internal consistency. For example, if a subject indicated in the exit interview that he had not changed his rank order of importance-for the attributes (safety and comfort), we would check this against the number of preference reversals the coder had identified. These solutions gave us sufficient confidence in the quality of the coding.

The method of analysis should not restrict our scope of attention beforehand (issue IV.4). This was the reason that we started with a very general model (Simon's model), paying the price of aggravating some of the previous validity issues.

It should be noted that our general model, while in our opinion valid for a range of problems, is by no means the only possible model. Especially within the area of problem solving there are many models that may form the basis of the design of a general model. It is worth considering the work of Jonassen (2000) that was pointed out to us after conclusion of our research While it deals with problem solving in an educational context, it is and is aimed not only at studying cognitive processes but also at designing tools for educating people in problem solving. As indicated earlier, an open mind when looking for possible theoretical foundations for this phase is essential.

5 The qualitative analysis (phase V)

The qualitative analysis was done according to the grounded theory approach. This approach will be described in Section 5.1, followed by the description of our analysis in Section 5.2. Validity issues are discussed in Section 5.3.

5.1 The philosophy behind the qualitative analysis: grounded theory

As discussed before, there are as yet no models available that describe importance-assessment processes. Therefore, a method was needed that imposed as few limitations in perspective as possible, so that the assessment process could be viewed with an open mind.

A well-known method that answers to these requirements is the grounded theory method (Strauss and Corbin, 1998). This method is meant to generate knowledge about concepts, and relationships between concepts, that are studied in qualitative research.

In this study, the sub-phases 1–5 as mentioned in section under phase V in Section 2 were followed largely, but not completely, sequentially.

5.2 The phases of the grounded theory approach in the qualitative analysis

5.2.1 Labelling phenomena

Identifying all attributes and the ways in which they were processed, as depicted in Figure 1, formed the start of the analysis (as related in Section 4.3). The first step in using the preliminary coding scheme for further qualitative analysis was identifying:

- the types of arguments used by the subjects (including the special category 'handling uncertainty')
- ways in which weights were linked to attributes and arguments, and the number of weight-attribute-argument combinations was progressively reduced.

As a preparatory step for further qualitative analysis in this phase, segments in the protocols that pointed at certain variables or activities were identified. These were subsequently inserted in the coding scheme. This is still a rather preliminary method, and rather subjective and explorative. But it shows a first structure hinting at phenomena. In order to give an impression, an example of a part of such an explorative qualitative coding scheme is given in Appendix 2. Essentially, this is the application of the coding scheme developed in phase III.

The next step was the identification of so-called 'working rules', i.e., general labels for the specific mental activities of the subjects. For example, subjects might say: 'I know that good suspension is important because my uncle had a minibus with poor suspension and that was very uncomfortable'. This would be labelled as 'use of personal experience'. Note that some mental activities may fall under more than one label, even though we tried to avoid this as much as possible. The grounded theory approach does not prescribe a set method for labelling phenomena, enabling a combination of creativity and systematic induction.

The key question here is, of course, under which conditions a subject's statement is considered for being re-formulated in more general terms in the first place. For some statements it is clear that labels are useful for describing the importance-assessment process. But especially statements that do not strike the researcher as relevant, if only because he has not encountered similar statements made by other subjects, are in danger of being unjustifiably ignored. That is why the phases of the grounded theory approach were gone through in an iterative fashion. All protocols were studied at least fifteen times in order to extract meaningful statements from them.

5.2.2 Discovering categories

After the working rules had been identified, they were grouped into categories. This was done by means of a number of iterations. First, the rules were simply grouped on the basis of their apparent similarity. For example, all rules that indicated a change in importance of a (sub) attribute were grouped under the category 'weight change rules',

and distinguished from rules that pertained to the elimination of (sub) attributes. Weight change rules and elimination rules pertain to an element of the problem space developed in Section 4, i.e., arguments for an attribute-weight combination.

Other categories that pertained to arguments could be identified as well, in particular the so-called reference source rules. These rules define the perspective which a subject takes while using an argument, for example the perspective of clients or of the drivers of the minibuses.

For some rules no suitable category could be found, so they were put into a rest category. An example is the wish to work as systematically as possible.

The next step was relating the categories to each other, sometimes aggregating categories. When the relationships were clear, this provided a system for naming the categories. This step will be dealt with in the next section.

5.2.3 Naming categories

Up till now, provisional names have been given to the categories. The next step was to name categories in terms of their place and function in the model representing the importance-assessment process. First, a general structure of the model had to be devised. The idea was to try and place categories in some logical sequence. This did not imply that subjects actually followed – or should follow – this sequence, but that in their thought processes certain phases could be distinguished. This approach is in line with Simon's (1979) problem-solving model (see Section 4) and with many other descriptive and prescriptive problem-solving models (Evans, 1991; Hicks, 1995; Lipschitz and Bar-Ilan, 1996).

With the categories that had been generated in the previous steps in mind, the protocols were examined, not at the level of individual statements, but now at a more general level of groups of statements pertaining to one and the same element of the problem space (see Section 4.2.1). The pattern that was discovered was that the subjects when performing the assignment applied four clusters of weighting activities:

- absolute weighting (see Section 4.2.2.2)
- relative dimensional weighting of sub-attributes pertaining to the same main attribute
- relative dimensional weighting of sub-attributes pertaining to different main attributes
- holistic weighting of the two main attributes safety and comfort against each other.

During the development of the model (see Section 4) we already identified these ways of weighting as legal moves. The qualitative analysis, however, showed that these hitherto isolated phenomena are essential in the importance-assignment process in terms of clustered activities, phases in Simon's terminology. This is an essential preparatory step towards measuring the significance of these clusters in a quantitative way. So, of all the phenomena identified in the preliminary analysis tool, only a few were necessary to form the backbone of our model. This was a vital gain from the qualitative analysis. Once the phases of the model were named, the categories discovered earlier were grouped under these phases whenever possible. Some categories that could not be linked to specific phases were classified as 'auxiliary activities' and analysed separately. The resulting model is presented in Heerkens and van der Heijden (2003).

5.2.4 Developing categories in terms of their properties and dimensions

While the analysis so far was focused on developing the model that describes the importance-assessment process, the next two steps were aimed at operationalising the model so that differences in thought processes between subjects within the framework of the model could be described, as a precursor to the quantitative analysis.

For the category 'processing of attributes', rather precise measurement was possible, as explained in Section 4.3. Several aspects of intermediate and final importance-judgements were also measured, like the number of preference reversals, the number of (sub) attributes eventually weighted and whether sub-attributes or main attributes were weighted in the final importance-judgement.

5.2.5 Establishing relationships between the various categories and with other variables

The most important relationships in our study pertain to the four clusters of weighting activities mentioned in Section 5.2.3. These formed the core of our model, which was completed with two structuring phases (see the distinction between problem structuring and problem solving in Section 4.2.2) and an evaluation phase found in many problem solving models).

Other relationships that were established are expressing dominant rules in terms of effort devoted to the clusters of activities in the phases as discovered. A working rule was dominant for a subject if it was observed both in the think-aloud protocol and in the exit interview protocol. We assume that, if this occurred, a working rule was really important for a subject, Furthermore, relationships with variables known from other theories on characteristics of expertise, characteristics of process planning and influences of bounded rationality were established.

This completes the description of the way the qualitative analysis was performed. The result of this analysis was a general model of the importance-assessment process, including some detailed descriptions of the way the individual subjects performed this process. Furthermore, it served as the basis for the quantitative analysis, which is discussed in the next section. As with the previous phases, we now assess how the validity issues described in Section 2 manifest themselves in the way we conducted the qualitative analysis.

5.3 Validity issues concerning the qualitative analysis

For this phase, the previously discussed issues under phase IV are again relevant, but they were already discussed adequately in Section 4.4 and we do not want to repeat ourselves. The only new issue is V.1–3. As for objectivity in establishing relationships; in this phase of our study, no definitive statements about relationships can be made; only qualitative arguments for relationships will be given. On the whole, in terms of Cooper and Schindler (2003), the research in this phase is more descriptive than causal. However, discussing perceived relationships with qualified colleague researchers to avoid 'tunnel vision' enhanced some objectivity. In a qualitative phase this is the best one can hope for.

6 The quantitative analysis (phase VI)

In Section 6.1, the way the quantitative analysis that was performed is described. In Section 6.2, some validity issues are discussed.

6.1 The steps in the quantitative analysis

The aim of the quantitative analysis was to answer the following questions:

- to what extent could the phases (clusters of activities) developed by means of the grounded theory approach in the previous section indeed be observed among a significant percentage of subjects? In other words, besides the fact that the phases are logically defendable, are they empirically relevant in a significant way as well?
- how much effort was devoted to each of the phases of the model of the importance-assessment process?

Counting of the properties and dimensions described in Section 5.3.4 was not part of the quantitative analysis. It was done during the course of the qualitative analysis.

The procedure for the quantitative analysis was relatively straightforward. Some salient aspects will now briefly be described.

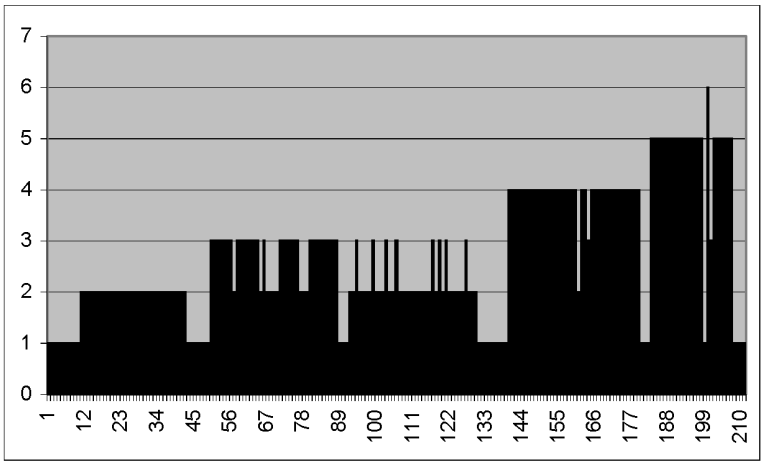
- *The development of a new coding scheme, independent of the scheme for the qualitative analysis*. The researchers determined how segments should be derived from the protocols and be allocated to the various coding categories. This turned out to be relatively simple. Segments are the parts into which the coding scheme is divided. They are the smallest entities in a protocol that can, thus, be coded independently of each other. A segment may correspond to a sentence, but it may also be part of a sentence, or even a single word (in our case, the naming of an attribute).
- *Development of the coding procedure*. The procedure developed was inspired by Chi (1997), Ericsson and Simon (1993), and van Someren et al. (1994). Two coders were asked to code all protocols. A focal point here was inter-coder reliability (Baarda and De Goede, 2001). Because the coding scheme comprised a large number of possible codes to assign, the probability of inconsistency seemed high. Furthermore, although the coding scheme seemed straightforward, the protocols were so diverse that it was to be expected that coders would sometimes have difficulties to distinguish between adjacent phenomena. Therefore, it was decided to have, in the beginning of the coding processes, a few meetings during which the coders could, along with the researcher, discuss possible problems. These meetings were not meant to directly adjust the coding but to improve the coding scheme. Before each meeting, each coder made a list of problems that he or she had encountered. Then, he or she re-formulated each problem in a general way. For example, if a coder hesitated between ‘evaluation’ and ‘attribute judgement’ in case a subject said: “If I give safety a weight of 0.8 and comfort 0.2, then that is too extreme. Safety is worth more than that”, then he or she would formulate as a general problem: “If a subject gives a judgement and then immediately modifies it, is the second judgement classified as a judgement or as an evaluation?” Then, during the meeting, it was decided that modified judgements would not count as evaluations

but as judgements. Afterwards, both coders would make any corrections necessary in the protocols. While this violates the rule that coders should work completely independent of each other, the advantages of improving the coding scheme 'on the fly' were considered greater than the drawbacks.

- *The coding itself*. First, the protocols were segmented, not by the researcher who had already done the qualitative analysis, but by independent coders. The numbers of segments associated with a model element of phase IV or a category of phenomena (phase as discussed in the previous section) were used for quantitative analysis. The results were stored in SPSS-files. SPSS, or Statistical Package for the Social Sciences is a widely used software program for the storage and analysis of quantitative data (Norušis, 1999).
- The analysis. The first step was to make frequency distributions of all phases as found in the qualitative analysis. Also, some specific issues were addressed that had caught the attention of the coders, like differences between male and female subjects. Also, it was assessed to what extent the subjects jumped back and forth between the phases, using a modified version of the approach of Lipschitz and Bar-Ilan (1996).

Typical examples of the types of results that our method of analysis can generate are as follows. The percentage of effort devoted to absolute sub-attribute weighting, measured in terms of the number of segments, is 27%. Fifty five percent of the subjects put some effort into heterogeneous sub-attribute weighting. For more details on this phase model we refer to Heerkens and van der Heijden (2003). Figure 2 shows graphically, for one particular subject, the phase to which each sequential segment belongs. If the subject had executed each phase sequentially in time (not jumping back and forth between phases) the figure would show six blocks of increasing height from left to right. This subject did not follow the phases sequentially. The analysis will be the subject of a future publication. We just want to illustrate the sort of results that the quantitative analysis can lead to.

Figure 2 Distribution of phases over sequential order of segments for one of the subjects



6.2 Validity of the quantitative analysis

The issues VI.1–2 pertain to this phase of the study. VI.1 concerns the content validity of the coding categories. For example, we used the number of segments associated with a particular phase as an indicator of the effort devoted to that phase. This was the most practical solution, but there are other indicators of effort, such as time. The content validity issue was all the more relevant because little research existed on importance-assessment processes. To determine the inter-rater reliability (issue VI.2) we used Cohen's Kappa. The value of 0.97 gave us confidence in the coding. We want to stress that the quantitative analysis was based on the results of the qualitative analysis, providing an extensive check on the logic of the operationalisation of the categories introduced and the objectivity of assumed relationships.

This completes the description of the research design. The next section is devoted to the assessment of the external validity of the research resulting from the chosen design.

7 Assessment of the external validity of the results (phase VII)

Up till now, we discussed the validity of separate phases in our research. This provides, in our view, an adequate picture of the internal validity of the research, which we think is, on the whole, adequate. What remains, then, is the external validity of the research. Note that we are not aiming at the external validity of our research method, which will be dealt with in the next section, but at the research itself. In our case, we used a very small sample (18 students), in a laboratory situation, not representative of the practical context in which the assignment which our subjects had to fulfill normally takes place. We deliberately designed an 'extreme case' so as to isolate the cognitive processes we wanted to study, yielding a richness of possible approaches that we feel could not be achieved in a 'real life' situation with actors constrained by their own experience and possible organisational pressures. So, we deliberately preferred internal validity and richness above external validity. With the model we developed, we are ready to look at real life situations where the richness of our model may not show (but then at least we know what is absent) and where perhaps new elements will be added. So, the external validity of our work is limited, but we accept that and are at least aware of it.

8 Discussion and recommendations for further research

Summarising, the research approach described here is new in several respects:

- it uses the think-aloud method for studying problem-solving of unstructured problems for which no accepted heuristics exist and the solution of which is highly determined by personal values
- the analysis of the think-aloud protocols has started with only a very broad model
- it concerns cognitive processes that, to our knowledge, have not been studied before, so no ready-made experimental set-ups were available.

In the illustration concerning our attribute weighting research project, we demonstrated that our design method enabled us to start without a clear theoretical frame of reference

and yet arrive at quantitative data. Usually, it is either one or the other. Therefore, we consider it to be a valuable addition to methods such as the grounded theory approach. The role of a tool for the preliminary analysis of was emphasised. This research method can be recommended for use in further research on importance-assessment processes, for example for studying experts, but also for other cognitive processes of a highly unstructured nature. The method is labour-intensive, but no other major practical problems were encountered. We feel that the method is especially worth considering for areas where psychology meets other disciplines, such as management studies, where methods like case studies and surveys are often used, and which provide very limited insight into cognitive processes. An example would be the process of strategy formulation in a company. In a case study, written material and interactions between actors can be studied, but what goes on in the minds of those actors can only partially be inferred from their actions. These cognitive processes may be highly personal and may not easily fit into models describing logical problem solving.

Yet, our research method could well be used in such a situation. Actors involved in strategy formulation processes could be given unstructured problems pertaining to various elements (phases) of the strategy formulation process and could be asked to solve them thinking aloud. No prior model of analysis would be required.

8.1 External validity of our research design scheme

The research we present in this paper was not aimed at developing a design, but at studying a certain type of cognitive process. The research design was merely a means to and not an end of our research, and the generalisation, induction if you want, of our research activities into the guidelines presented in this contribution are a fall-out of our research, an added bonus. Consequently, the general validity of the design has not been determined. We can merely present the requirements to assure validity the way we believe these requirements were fulfilled on our research. We do not see this as a major drawback, however. Every generic design needs to be adapted to whatever research project it is used for, so the validity needs to be assessed in every specific case. We believe that the validity requirements we define make this possible. All in all, we believe that our design is both generic enough and yet addresses the peculiarities of research on cognitive processes sufficiently, to be used for a wide range of research on cognitive processes, as long as the validity requirements stated in Section 3.1 are evaluated at the beginning of each specific research project.

8.2 Further research

Most problems that we encountered in our research were specific for our research problem. The choice of subjects and the nature of the attributes to be weighed (not too straightforward, but not too difficult either), and the way of providing the subjects with enough information to conduct the assignment without overloading them required long deliberations. The only significant problems not specific to our research but typical for our research design scheme pertained to phase IV: the design of a general model for the preliminary data analysis. The first challenge was to assure that there was indeed no model specifically aimed at importance-assessment processes. Where do you look, and when can you be sure to have looked in all likely places? The next challenge was to find the clues for the very first outlines of the general model. How can you be

specific enough to make the model have added value without narrowing your view too early so that you ignore more promising paths? The following experience-based tips may be useful for researchers who want to use our design scheme:

- First write down your own ideas about what the general model could look like on the basis of the protocols, and only then start surveying the literature for more ideas. Starting to examine the literature too early may 'force' you in specific lines of thinking too early. You have to have a broad arsenal of possible approaches before you settle for a particular one. Remember that it was the lack of usefulness of the literature that drove you into our research design scheme in the first place.
- Be creative, play 'advocate of the devil', do not dismiss any idea out of hand and look at the reality of your own daily life for inspiration. One of the authors used writing music, fiction and articles for an aviation trade journal as inspiration for the model.
- Re-read the protocols, not necessarily systematically, but use your intuition when deciding which (parts of) protocols you want to examine. Maximise serendipity by trying to translate everything you encounter, even, or especially, not related to your research, in terms of the problem that is so be modelled. The inspiration for the notion of the importance of absolute weighting came from a poem that went with a Christmas present for one of the authors.
- Use as many approaches to your problem as possible, and only choose a definitive approach (general structure of your model) when you feel you have run out of new ideas.
- Before you choose your definitive approach, think (and read) about possible ways of modelling. Miles and Haberman (1994) may provide inspiration for qualitative modelling of complex processes.
- When having your own ideas more or less in place, talk about the subject with others. Do not be alarmed if they have different views on the model to be designed; in this stage, nothing is settled yet. Be prepared that people will tell you that your problem was solved a long time ago, let them explain to you their models, which will likely be of little use, thank them for their ideas and use the sensible elements in your own model.

These tips may not appeal to everyone, but when no theoretical framework exists for your research, you have to grab every possibility to get 'method in the madness'. The actions described in the tips were of great help to us. More tips may be found in the literature about creativity and problem solving.

Our conclusion is that our method provides a way to study individual cognitive processes while introducing an organisational setting. In this sense, the method has the potential of offering the best of both worlds. Moreover, this study can provide the basis for further research in three areas:

- Applying the method of analysis to importance-assessment processes in other contexts. The subjects in this study were laymen. It has yet to be proved that the concepts used can describe importance-assessment processes as conducted by experts, or by actors working in a real organisational context. Actors involved in the acquisition process of, for example, a capital goods, within an organisation could be

asked to participate in an experiment similar to the one we have conducted in our research. The assignment would of course have to be adjusted, that is to say, the company, the capital good and attributes to be weighted would have to be in line with the real-life acquisition process that the actors are involved in.

- Development of similar frameworks of for preliminary analysis for other types of cognitive processes concerning the solving of complex (unstructured) problems. An area that comes to mind is problems where both rational reasoning and values are involved (as is the case with importance-assessments), like business decisions where ethics plays a role. For example, research has been done on the choice of methods for transporting and storing dangerous (radio-active) materials. See, for example, Keeney, 1992). While elicitation methods exist for assessing the perceived importance of attributes relevant for the choice, it would be interesting to study the way actors handle the ethical problem of weighting the importance of subjecting people to safety hazards vs., for example, financial attributes.

References

Baarda, D.B. and De Goede, M.P.M. (2001) *Basisboek Methoden en Technieken*, Stenfert Kroese, Groningen.

Blalock, H.M. (1981) *Social Statistics*, Revised 2nd ed., McGraw-Hill, Auckland.

Borcherding, K., Schmeer, S. and Weber, M. (1995) 'Biases in multi-attribute weight elicitation', in Caverni, J.P., Bar-Hillel, M. and Jungemann, H. (Eds.): *Contributions to Decision Making-1*, Elsevier Science BV, Amsterdam, pp.3–29.

Chi, M.T.H, Glaser, R. and Farr, M.J. (Ed.) (1988) *The Nature of Expertise*, Lawrence Erlbaum Associates, Hillstate.

Chi, M.T.H. (1997) 'Quantifying qualitative analyses of verbal data: a practical guide', *The Journal of the Learning Sciences*, Vol. 6, pp.271–315.

Cooper, D.R. and Schindler, P.S. (2003) *Business Research Methods*, 8th ed., Irwin/ McGraw-Hill, Jaartal Controleren, Boston.

Davison, G., Vogel, R. and Coffman, S.G. (1997) 'Think aloud approaches to cognitive assessment and the articulated thoughts in simulated situations paradigm', *Journal of Consulting and Clinical Psychology*, Vol. 65, pp.950–958.

de Leeuw, A.C.J. (1997) *Organisaties; Management, Analyse, Ontwerp en Verandering*, Van Gorcum, Assen.

Easterby-Smith, M., Thorpe, R. and Lowe, A. (1993) *Management Research, An Introduction*, Sage Publications, London.

Eisenhardt, K. (1989) 'Building theories from case study research', *Academy of Management Review*, Vol. 14, pp.532–550.

Ericsson, K.A. and Hastie, R. (1994) 'Contemporary approaches to the study of thinking and problem solving', in Sternberg, R.J. (Ed.): *Thinking and Problem Solving*, Academic Press, San Diego, pp.37–79.

Ericsson, K.A. and Simon, H.A. (1993) *Protocol Analysis: Verbal Reports as Data*, 2nd ed., Books/MIT Press, Cambridge, MA.

Ericsson, K.A. and Smith, J. (1991) *Towards a General Theory of Expertise*, Cambridge University Press, Cambridge.

Evans, J.R. (1991) *Creative Thinking in the Decision and Management Sciences*, South-Western Publishing Co, Cincinnati.

Harris, B. (1992) 'Natural translation: a reply to Hans P. Krings', *Target*, Vol. 4, pp.97–103.

Harte, J.M and Koele, P. (1995) 'A comparison of different methods for the elicitation of attribute weights: structural modeling, process tracing and self-reports', *Organisational Behavior and Human Decision Processes*, Vol. 64, pp.49–64.

Heerkens, J.M.G. and van der Heijden, B.I.J.M. (2003) 'On importance assessment and expertise in non-routine decisions; an exploratory study on the cognition of weighting processes of capital goods' attributes,' *International Journal of Management and Decision Making*, Vol. 3, pp.370–398.

Hicks, M.J. (1995) *Problem Solving in Business and Management; Hard, Soft and Creative Approaches*, International Thomson Business Press, London.

Hunt, E. (1994) 'Problem solving', in Sternberg, R.J. (Ed.): *Thinking and Problem Solving*, Academic Press, San Diego, pp.215–232.

Jaccard, J., Brinberg, D. and Ackerman, L.J. (1986) 'Assessing attribute importance', *Journal of Cvonsumer Research*, Vol. 12, pp.463–467.

Jonassen, D.H. (2000) 'Toward a design theory of problem solving', *Educational Technology Research and Development*, Vol. 48, No. 4, pp.63–85.

Kahnemann, D. and Tversky, A. (2000) (Ed.) *Choices, Values and Frames*, Cambridge University Press, Cambridge.

Keeney, R.L (1992) *Value-focused Thinking*, Harvard University Press, Cambridge MA.

Keeney, R.L and Raiffa, H. (1976) *Decisions with Multiple Objectives, Preferences and Value Tradeoffs*, John Wiley and Sons, New York.

Lipschitz, R. and Bar-Ilan, O. (1996) 'How problems are solved; reconsidering the phase theorem', *Organisational Behavior and Human Decision Processes*, Vol. 65, pp.48–60.

Miles, M.B. and Haberman, A.M. (1994) *Qualitative Data Analysis; An Expanded Sourcebook*, Sage, London.

Newell, A. and Simon, H.A. (1972) *Human Problem Solving*, Prentice-Hall, Englewood Cliffs, NJ.

Norušis, M.J. (1999) *SPSS 9.0; Guide to Data Analysis*, Prentice Hall, London.

Póyhónen, M. and Hämäläinen, R. (1998) 'Notes on the weighting biases in value trees', *Journal of Behavioral Decision Making*, Vol. 11, pp.139–150.

Saaty, T.L. (1980) *The Analytic Hierachy Process*, McGraw-Hill, New York.

Séguinot, C. (1996) Some thoughts about think-aloud protocols', *Target*, Vol. 8, pp.75–95.

Simon, H.A (1979) 'Information processing theory of human problem solving', in Estes, W. (Ed.): *Handbook of Learning and Cognitive Processes*, Lawrence Erlbaum Associates Inc., Hillsdale, Vol. 5, pp.271–295.

Strauss, A. and Corbin, J. (1998) *Basics of Qualitative Research; Grounded Theory, Procedures and Techniques*, Sage Publications, Thousand Oaks CA.

Swanborn, P.G. (1973) *Variabelen en Hun Meting*, Boom, Meppel.

Swanborn, P.G. (1987) *Methoden van Sociaal-wetenschappelijk Onderzoek*, Boom, Meppel.

Timmermans, D. (1993) 'The impact of task complexity on information use in multi-attribute decision making', *Journal of Behavioral Decision Making*, Vol. 6, pp.95–111.

van Riemsdijk, M.J. (1999) 'Dilemma's in de bedrijfskundige wetenschap; enkele achtergronden', in van Riemsdijk, M.J. (Ed.): *Dilemma's in de Bedrijfskundige Wetenschap*, van Gorcum, Assen, pp.1–27.

van Someren, M.W., Barnard, Y.F. and Sandberg, J.A.C. (1994) *The Think-aloud Method; A Practical Guide to Modelling Cognitive Processes*, Academic Press, London.

Verschuren, P. (1980) *Strukturele Modellen Met Theoretische Variabelen*, PhD Thesis, University of Nijmegen, Nijmegen.

Whitney, P. and Budd, D. (1996) 'Think-aloud protocols and the study of comprehension', *Discourse Processes*, Vol. 21, pp.341–351.

Appendix 1: Example of an attribute-processing scheme

The schemes should be read as follows. Safety always gets the number 1 and comfort number 2. Decomposed attributes at the first level get the numbers 1.1, 1.2, 2.1, 2.2 etc. At the second level, the numbers consist of three digits and can be, for example, 1.1.1, 1.1.2 etc. A letter placed after a certain attribute number means that the attribute is a re-formulation. If an attribute is the abstraction of another attribute, this is noted between brackets. An integration is always the result of two of more attributes being processed, and is also indicated between brackets.

The sub-attributes of the first level are listed as much as possible in the order in which the subjects mentioned them.

The processing of 'safety' by subject 4

1 Safety

- 1a If an accident happens, you want to get out in one piece, preferably unhurt (specification)
- 1b I want to get out in one piece or with very minor injuries, but not so that I can sit in a wheeled chair for the rest of my life
- 1c Accidents

1.1 Number of deaths per year with a certain brand

- 1.1a Number of accidents with which it has occurred
- 1.1b Number of deaths per year with accidents (from the context it is clear that it is meant per type)
- 1.1c Accident numbers (from the context it is clear that it concerns deaths per year)
- 1.1d Maximum so many deaths per year
- 1.1e How many deaths per year with accidents and with how many accidents does this happen?
- 1.1f Number of deaths per year

1.2 Number of serious injuries

- 1.2a Number of serious injuries per year per accident
- 1.2b How often does it occur (serious injuries)
- 1.2c Figures about serious injuries
- 1.2d Number of serious injuries per year
  - 1.2.1.1 Paralysed (downwards) from a certain body part or really loose a body part
  - 1.2.1.2 Paralysed
  - 1.2.1.3 Body part coming off

1.3 Seatbelts

1.3a Are seatbelts in the car?

1.4 Seat broke loose

1.5 Anti-skid system

1.6 Are there headrests?

1.6a Headrests

1.7 Can headrests be adapted?

1.7a Are headrests adaptable?

1.7b Are they adjustable in height (no specification because this is what he meant with 1.7 and 1.7a)

1.7c Are headrests adjustable?

1.7d Adjustable headrests

1.8 Safety for driver

1.9 Safety for assistant-driver

1.10 Safety for passengers

1.11 Airbag

1.11.1 Airbags on the side

1.12 How does a bus fare if you smash into it from the front, the rear, and the side and from above?

1.12a With crash tests what was the result (abstraction)

1.12b Result with type of accident

1.12c Result with crash tests

1.12.1 If an airplane crashes on your car

1.12.2 If such traffic pole like you have in Enschede comes crashing into your car from underneath

1.12.3 From the side they come

1.12.3a If someone comes from the side

1.13 To what extent does a baby sit safely in the car?

1.13.3 Has it got baby seats?

1.13.4 Does the possibility exist to install them (baby seats)

1/13.1/1.13.2a Baby seats: are they there, can they be installed (integration)

1.13.3 Do baby seats have to be with the face forward or with the face rearward?

1.13.3a Which baby seats are dangerous, which are not dangerous?

Appendix 2: Example of the qualitative coding scheme

Remarks of the coders are made in italics or indicated by the word 'note'.

The numbers in the right column refer to 'working rules' that were identified during the qualitative analysis (see Section 5.3.1).

The distinction between decision rules, weighting rules, implicit and explicit rules could not be unambiguously maintained and was not used during the analysis.

	Subject 11
Number of reformulations safety	0
Number of reformulations comfort	4
Reformulation safety taken into account, unity plus level of measurement	Original formulation (safety) taken. Level of measurement at least ordinal
Scale?	No, so no common scale
Reformulation comfort taken into accounts, unity plus level of measurement	Original formulation (safety) taken. Level of measurement at least ordinal
Scale?	No, so no common scale
Number of decompositions/specifications safety plus levels	First level: 4, second level: 1 specification
Number of decompositions/specifications comfort plus levels	First level: 7, second level: 9, third level: 1 specification, fourth level: 2
Number of sub-attributes safety taken into account plus level of measurement	0
Scale?	Not applicable
Number of reformulations taken into account sub-attribute safety	Not applicable
Number of sub-attributes comfort taken into account, plus level of measurement	0
Scale?	Not applicable
Number of reformulations taken into account sub-attribute comfort	Not applicable
Number of integrations/abstractions sub-attributes safety	0
Number of integrated/abstracted sub-attributes safety taken into account, plus level of measurement	Not applicable
Scale?	Not applicable

	Subject 11
Number of integrations/abstractions sub-attributes comfort	'You are picked up at your door and dropped at Schiphol' abstracted to 'comfort of the whole journey' and 'you will be waited for if you have a delay of less than 1.5 hours' abstracted to ' you do not have to adjust to train hours, but Plane Drive will adjust to you'
Number of integrated/abstracted sub-attributes comfort taken into account, plus level of measurement	0
Scale?	Not applicable
Number of (sub)-attributes that explicitly fall off after being taken into account	0
Weights of level of measurement	Interval
Safety weight given for every score (sub) attribute?	One weight for all scores
Safety weight given for every score (sub) attribute comfort?	Weight independent of comfort score
Comfort weight given for every score (sub) attribute?	One weight for all scores
Comfort weight given for every score (sub) attribute safety?	Weight independent of safety score
Level of conjunction limit given[1]	No conjunction limit
Common denominator?	No
Preference reversal	Number of abstractions weighted: 2
	Gross: 2
	Net: 0
Decision rules explicit[2]	
System in decomposition etc.	Decomposition in comfort for passengers and comfort for driver 2.2
Conjunction/rules for exclusion/elimination	Decomposition in comfort for the bus and comfort for the whole journey 2.2
Rules for scale construction	You have to give a weight factor to sub-attributes. Other test persons are indeed doing this, but are not addressing it explicitly as rule/goal and*: she only compares sub-attributes of the same main attribute*
Rules for different scores	With some sub-attributes: pairs wise comparison. *She indicates that this is the method to get weights of sub-attributes, though she doesnot mention AHP, but she does not do the method* 3.0.6.2
Patterns information gathering	Safety and comfort are close together: *is this about weight or about empirical relation?*
	Company is telling things in brochure as they think they matter.[3]

	Subject 11
Global level of measurement Etc.	
Decision rules implicit	Level of measurement sub-attributes as known almost always ordinal or yes/no, 1 time nominal, 1 time ratio
For categories see decision rules explicit	First invent sub-attributes yourself, then brochure 2.3.3
	'Weight' is interest. 5.1
	Weights normalised to 1. 3.0.9
Weighting rules explicit[2]	Clients' perspective: comfortable sitting is important for the passenger (weight) 3.0.5.3.2
External reference frame	But with that one over there I see those benches, I would not sit on them on the way to Schiphol 3.0.5.1.1
Thought experiments, procedure as well as evaluation of weights, for example by hypothetical scores	For the chauffeurs comfort devices like cruise control are important[4].
Restricted number of values of weights?	The company itself emphasises comfort. 3.0.5.5.3
Systematic pairs wise comparison Etc	I would think seatbelts are more important than an airbag. An airbag only goes for the driver and the passenger sitting next to him 3.0.4.10
	Car frame is important for the whole car, so that has to stand above seatbelts. The interview shows that she means, that the whole car is protected. 3.0.4.7 and 3.0.4.10
	… For passenger's comfort and driver's comfort, because that are two different things according to me. And because driving people is your business, the customer's comfort is most important, and that is what you have got to look at mostly. 3.0.4.10
	Yes, of course comfort is not only characteristics of the car, (…) the whole journey is comfortable, then comfort also contains being picked up at your door and dropped at Schiphol, (…) yes I,m persuming here it's about those cars.[5]
	Comfort at least equals train, now it seems more to me, because you are picked up at your door. 3.0.5.9.1
	Safety is really important, but I think comfort is even more important, because you already assume some sort of safety at all vans. Later she says: there are not unsafe cars anymore. 3.0.4.1
	Comfort is the that the passenger noticed most.[6]
Weighting rules implicit	Weights in last instance determined by test person and customer
For categories see weighting rules explicit	3.0.5.10.1 and 3.0.5.10.2
Handling uncertainty	If something happens one time, safety is really important.[7]
Chance on event as attribute	The chance for an accident is of course really small, but if it happens, you have to get a van that is prepared. 4.3
Other weight	

	Subject 11
Extra attribute	
Multiple weights	
Discount on score	
Rules mentioned during interview[8]	Yes, if you read through the assignment, just what is coming in your head first about safety and comfort. (…). And then I checked safety and comfort at both cars and if there were any aspects I did not mention.[9]
Rules mentioned during interview[8]	Yes and with comfort I've mostly let myself lead by passengers comfort: Not clear what she means: looking at protocol, I do not assume she is for example leaving chauffeurs comfort out of consideration. Later she also says: I also mentioned the chauffeur's comfort. 3.0.5.3.3
	Well at the VW-van, I would not want to sit in there at such a bench the whole journey to Schiphol.[10]
	Then I thought comfort could be seen as do not having to go to the station to take the train and to adjust to certain hours, but as comfort of having a company picking you up at your door. (…) Especially with that aspect of comfort, that you can see much wider than just the car. But because you have to weigh between those vans, it appears to me that you still have to look at the car and not at the rest. 3.1.1.1
	But that (driver's comfort) I would find less important than comfort of passengers you have to accommodate, because they are your customers. 3.0.4.10
	So, that way I've made that assessment and that (driver's comfort) seen as less important and did not mention it the driver's comfort anymore. 3.0.4.10 and 3.1.1.4
	Comfort is that which the passenger will see and feel, and statistically speaking chances of an accident are rather slim. 4.3
	And because the car frame is for the whole car and for all passengers, I thought that was more important than just airbags that are only for the people in the front seats. Or you have to have airbags everywhere, but still then I would find the car frame for the whole car most important, because it is the outside, and if you have an accident, that still is what gets damaged where it gets hit by forces. 3.0.4.10 and 3.0.4.7
	And because I think safety is a bit more important (because if you need it you've got to have it) I settled for a bit across the middle. But not much, because I think that passenger comfort is quite important. 4.1 and 3.0.7.2
	Weights in last instance determined by subjects and clients: 3.0.5.10.1 and 3.0.5.10.2
What kind of information should you have (afterwards)?	None
What will be done differently next time?	Entangle even more sub-attributes with comfort
	Reading material a little bit better (to be able to entangle more sub-attributes)

	Subject 11
	Reading the assignment better, so it's clear that it's about the van's comfort and not about the comfort of the whole journey
	She would rank sub-attributes and assign rates to main attributes the same way next time
Remarks	During the interview she mentioned that, not knowing if it was about comfort of the van or comfort of the whole journey, was a moment of trouble. But it appeared from protocol this did not affect the method

[1]Even when sub-attributes are not weighted, they can be the conjunction limit indicator.

[2]Indicate if it is about alternatives, attributes or weights.

[3]implication: do not just acknowledge the folder of Plane Drive 3.0.10.4.

[4]I am taking that she means the chauffeur's opinion (she also said 'from the chauffeurs point of view'), although it can be that only she thinks it is important. 3.0.5.4.2.

[5]The interview shows that she leaves attributes, that are not considering vans but the whole journey, out of consideration. 3.1.1.1.

[6]Not very clear. 3.0.5.3.1.

[7]I assume because of (big) consequences. 4.1

[8] with analysis: do not consider rules that are mentioned during interview, but are not traceable to think aloud session. Rules mentioned after probing by the researcher are not considered: chance of rationalisation afterwards is too big.

[9] First make something up yourself, then brochure during decomposition. 2.3.3.

[10] Own perspective, score. 3.0.5.1.1.

Appendix 3

A checklist for analysing cognitive processes using exploratory think-aloud experiments:

This checklist is meant as help, not as 'a harness'. Eventually, the validity criteria as discussed in the paper are the ultimate test of how well our method has been applied. But we strongly urge the reader to at least take note of the points mentioned in our checklist.

We will not go deeply into issues that are of concern with research in general, like the organisation of the think-aloud sessions.

Phase 1: determining the research method

1.1 Determine what has priority: internal or external validity. Some issues to consider:

- 1.1.1 If little is known about a cognitive process, it may be best to study it in isolation, in order to determine which variables are involved.
- 1.1.2 If it is not clear which variables may influence the process, then it may be best to study the process in an environment that gives maximum freedom to the subjects (not hindered by, for example, organisational pressures), so that as many manifestations of the process can be observed but the variables that are part (indicators) of the process can still be distinguished from variables influencing it.

1.2 Establish the character of your problem statement. Is it suitable for a think-aloud experiment?

- 1.2.1 Is the problem statement about individual actors or groups?
- 1.2.2 Do you want to know the subjects' concurrent thoughts? Or do you want their opinions, the results of their thoughts, or thoughts of the past? Use the think-aloud method only in the first case.
- 1.2.3 Can you isolate the subjects (in a room with an experimenter and a recorder)?
- 1.2.4 Can you give the subjects an assignment that induces the processes you want to study (the subjects need to think while solving the assignment, not tell you what they thought when they solved the assignment in the past)?
- 1.2.5 Do the subjects have the information they need without having to absorb so much new information that the cognitive processes to be studied drown in information processing activities?
- 1.2.6 Do you have a clear enough idea of the nature of the processes that you want to study so that you can feel confident that you can design an assignment that induces these processes (even though you may not know which variables to study)?
- 1.2.7 Can you give the subjects an assignment that is 'new' enough to provoke mental processes (instead of memories of past processes or automated behaviour) and yet is not so difficult that the subjects know nothing to say?

1.3 Determine whether you have the resources available for a think-aloud experiment.

- 1.3.1 Space for conducting the experiment (sometimes all sessions have to be done at the same time so as to prevent interaction between subjects).
- 1.3.2 Recording equipment etc.
- 1.3.3 Personnel for making protocols (transcripts) of the sessions. It is best that the researcher does not do that him or her self, so as to be able to study the protocols with an open mind.
- 1.3.4 Time to ponder over the protocols over an extended period, in order to grasp the meaning of the unstructured data.

Phase 2: designing an experiment

2.1 Design the assignment. Consider the following issues:

- 2.1.1 The general problem area. Do the subjects have to be familiar with it or not (how non-routine does it have to be)? Three possibilities:
  - the subjects have experience with the problem area in their daily private or working life (i.e., buying food)
  - the subjects have no experience with the problem area from their own experience but are likely to have read or heard about it and can connect it to their own world (i.e. buying a minibus)
  - the subjects have no experience with the problem area and it is far removed from their own world (i.e., buying an aircraft).

2.1.2 The complexity of the assignment. If there is no reason to suppose that the complexity influences key variables, reduce complexity vigorously, so as to facilitate verbalisation, but keep the problem 'new' enough to discourage automated behaviour. For example: the structure of a cognitive process may not be influenced by complexity, only the relative importance of elements of the structure.

2.1.3 If information is to be supplied to the subjects, determine whether 'real-life' information is suitable or whether artificial information has to be designed. The latter can be more customised but may be less believable and motivating. If 'real-life' information is concise enough, it may be preferred.

2.1.4 Identify (during practice sessions) whether the assignment is clear. When you do not know the exact nature of the cognitive processes to be studied, you may not foresee erratic behaviour induced by the assignment. In our case subjects sometimes compared scores on attributes instead of their relative importance.

2.1.5 Sometimes the subjects should do more than you want to study. In our case, we let the subjects give weights to attributes, just to find out whether different subjects gave different weights. If they had all given the same weights the assignment could have been biased.

2.2 Other issues concerning the designing of a think-aloud experiment are not specific to the exploratory research on cognitive processes here and we will not cover them. Examples discussed in the paper are: avoiding external disturbances, and the artificiality of the task (bearing points 1.1 and 1.2.3 in mind).

2.3 Also, we suggest consulting someone with experience in these kinds of experiments. It is no use re-inventing the wheel and becoming an expert in think-aloud experiments if the experiment is merely a means to an aim: becoming an expert in certain cognitive processes.

Phase 3: designing a data collection method

3.1 Points 2.2 and 2.3 apply here.

Phase 4: designing a tool for preliminary analysis

4.1 Before conducting the actual experiment.

4.1.1 Consider which variables, that is, mental activities, and characteristics of the goal of the process (the problem to be solved) may be of any relevance. Try to be as comprehensive as possible.

- Use free association or brainstorming techniques, analogy reasoning, creativity, and conversations with both experts and lay-men etc. It is best not to study literature yet: it may constrain your creativity
- Use serendipity: look at the world from the perspective of your research and try to see traces of this perspective in real life. It may give inspiration.

4.1.2 Then search the literature for:

- More variables etc. that could be relevant.
- Definitions and conceptualisations of relevant variables, so as to get deeper insight in them.
- Possible models that show relationships between the identified variables (at the conceptual, at the interdependence (input/output) or at the statistical/causal level). The model presented in Figure 1 in Section 4.2.1 relates variables at the conceptual level; the formula in the same section shows a causal relationship. Note that sometimes, models from en entirely different discipline can provide valuable inspiration.

These actions will enhance construct validity.

4.1.3 If preferred, consult a book about possible forms of (visual) presentations of concepts, as a source of inspiration.

4.1.4 Design the preliminary model.

4.1.5 Identify the elements in the preliminary model, insofar as they were not identified before designing the model.

4.1.6 Define the various (mental) actions the subjects could take in order to process the various elements of the model. In our case: for example, processing attributes and formulating arguments for assigned weights.

4.1.7 Consider the various ways in which the elements and (mental) actions could possibly manifest themselves in a verbal protocol. Pay special attention to distinguishing elements from each other (so as to enhance coder consistency and content validity). This is your tool for preliminary analysis.

4.1.8 'Fill in' the tool with a few practical examples, preferably from a subject that you feel very attached to, so as to see if the took fits real-life situations.

4.2 Conduct the experiment.

4.3 After the experiment and the typing-out of the protocol:

4.3.1 Read the protocols with an open mind, without thinking about the model you made in 4.1.4. Make notes of everything that strikes you, because after a few times you will not see the striking features anymore. Repeat these step five times or so in a number of weeks.

4.3.2 Check whether striking features seen in one protocol resurface in any form also in other protocols.

4.3.3 Compare your observations so far with the model and adjust the analysis tool (4.1.7). Glance over steps 4.1.1 and 4.1.2 to see whether any modifications need to be made.

4.3.4 Design a simple system of identifying the elements of the analysis tool in the protocol, like underlining etc. You can use several systems in parallel. Examples can be found in Appendices 1 and 2.

4.3.5 When reading through the protocols, check your own consistency during coding. Experience may cause you to read the last protocol with different eyes from the first. Keep refining your way of coding. If you see a new phenomenon, check whether it was present in protocols you coded earlier.

Phase 5: the qualitative analysis

We will address this phase only briefly since the Grounded Theory approach that is the core of this phase has been well described elsewhere (see the references).

5.1 Name the various elements, (mental) actions and striking features (from now on to be called phenomena) found in the protocols. Note: it will sometimes be difficult to decide whether two phenomena are or are not manifestations of the same higher-order phenomenon. By the time you have seen the protocols so often that you sincerely hate them, these and other dilemmas will be resolved on way or another. NB: a name need not mean a single word; it may also be a short description. In our research, we could describe most phenomena as working rules.

5.2 Discover categories, i.e., group wise similarities and differences between phenomena. In our case: for example, working rules considering the processing of attributes or the formulation of arguments for weights.

5.3 Group the phenomena in categories. Doing this consistently requires a lot of iteration between this step and the previous two.

5.4 Describe categories in their properties and dimensions, so that phenomena can be characterised as belonging to a certain category. Whereas steps 2 and 3 were essentially logical induction, this step is systematic deduction. Ideally, phenomena discovered in completely new protocols can be grouped reliably into categories, based on their characteristics, even without knowledge of the logic of steps 2 and 3.

5.5 Establish relationships (statistical, causal, sequential etc.) between the categories. This step yielded our seven-phase Weight Assessment Model.

5.6 Design a way to present the categories and their relationships (in a sequential listing, a model, a diagram etc.).

5.7 Check whether all elements of your preliminary analysis tool were addressed adequately during the qualitative analysis.

5.8 It may turn out to be necessary to modify the tool for preliminary analysis during this phase, and adjust (part of) the qualitative analysis accordingly.

Phase 6: the quantitative analysis

This step will also be addressed only briefly, since it does not differ much from standard analysis procedures.

6.1 Define the purpose of the quantitative analysis, so that it is possible to:

6.2 Define the phenomena to be looked for in the protocols.

6.3 Develop the coding scheme with maximum content validity. This scheme may well not be as exact as in most quantitative research, so:

6.4 Decide how to handle inter-coder reliability and concurrent development of the coding scheme. This is especially important if the number of protocols is limited, so that the processes of coders gaining experience may cover a large part of the protocols. See Section 6.6.1 for our solution.

6.5 Develop an analysis procedure.

6.6 Perform the coding and analysis.

6.7 Interpret the results.

Phase 7: establishing validity

There are no differences with standard procedures here.