On Importance Assessment and Expertise in Non-Routine Decisions

An Exploratory Study on the Cognition of Weighing Process of Capital Goods' Attributes

Hans Heerkens, School of Management, Public Administration and Technology, Department Operational Methods and Systems Theory, University of Twente, PO Box 217, 7500 AE Enschede, the Netherlands, e-mail: j.m.g.heerkens@sms.utwente.nl.

Beatrice van der Heijden, School for Management, Public Administration and Technology, Department Human Resource Management, University of Twente, PO Box 217, 7500 AE Enschede, the Netherlands, e-mail: b.i.j.m.vanderHeijden@sms.utwente.nl.

Abstract

How do actors involved in the acquisition of a capital good assess the importance of its attributes? And what is the role of expertise? Numerous instruments exist for measuring the importance attached to attributes, but little is known about the importance assessment process that precedes these importance judgments.

Expectations concerning the behaviour of actors facing non-routine importance assessment problems are tested, yielding some interesting results. Firstly, the behaviour of these actors is consistent with a newly developed phase model. Even with a non-routine problem, structuring the assessment problem takes less effort than the actual weighting. Surprisingly, weighting attributes in isolation gets much more emphasis than weighting them against each other, despite the latter being the essence of importance judgments. Despite the subjects being laymen, they showed high confidence in their work.

Finally, predictions concerning the behaviour of experts are made, based on Van der Heijden's [1] dimensions of expertise.

Keywords

Attribute weighting, decision theory, expertise, importance, utility.

Biographical notes

Hans Heerkens is assistant professor in the field of research methodology and problem solving. He also teaches on strategic management in the aerospace industry.

Beatrice van der Heijden is associate professor in the department Human Resource Management. She is specialised in career development, ageing and employability research. Moreover, she does some consultancy work in the field of social policy.

1: Introduction

Let us first introduce the problem area with a brief case description. Consider a firm that transports passengers to and from an airport some 150 km away. Its clients are, for various reasons, unable or unwilling to travel by car or by public transport. Years ago, a previous manager bought a fleet of Volkswagen minibuses that have to be replaced in the coming year. The minibuses that are available on the market differ from each other in many respects, such as ease of maintenance, running costs, safety features, passenger capacity, passenger comfort and so on. These different minibuses' characteristics can be labelled with the term 'attributes.'

As the present, recently employed, manager wants to standardize on one type of minibus, he has to decide which type to acquire. In order to do this, he needs to decide how important the various characteristics of the minibuses are to him. Is ease of maintenance more important than passenger comfort? If this is the case, and if he has to choose between minibus A that is not so comfortable but easy to maintain and bus B that is more comfortable but also more difficult to maintain, he'll probably choose bus A. But what if bus A is somewhat easier to maintain, but bus B is a lot more comfortable?

Then the manager has to decide how much more important ease of maintenance is to him compared with comfort.

This example has several characteristics that managers are often confronted with. Firstly, his challenge is to make a choice in a situation where it is unlikely that one alternative is the most favourable with regard to all characteristics. It cannot be expected that one minibus is best in all respects compared to all other minibuses. Secondly, the task at hand is not a routine decision. It will occur no more often than once every so many years. Besides, on each occasion the attributes of the available minibuses will probably have changed, as may be the case regarding the importance of these attributes. Minibuses may have become more comfortable, but clients may rate comfort increasingly important. The manager can be considered an expert when running a fleet of minibuses is concerned. But to what extent does this specific expertise help him?

A rational manager who faces a non-routine decision like the one in the case description basically has two tasks:

1. He has to decide on the relevant attributes and on the importance of each of these attributes, in terms of its influence on the final choice. The thinking process about attribute importance is called the process of *importance assessment* and results in an *importance judgment*: the weight given to an attribute.
2. He has to assess the *values of the attributes* of each alternative to be taken into consideration. Given the importance judgment on each attribute, the most attractive alternative can then be chosen.

In this article, we will focus on the first problem: how do actors in non-routine decision-making arrive at an importance judgment? The goal of this study is firstly to *develop a generic model of the importance assessment process* and secondly to *identify the role of expertise in non-routine decision-making*.

From a managerial perspective, the importance assessment problem is of interest, as it neither implies a completely logic-driven problem nor a completely value-driven problem. The concept of 'importance' implies that the actor concerned has personal goals, norms and values, or at least preferences with an affective component. But no actor, at least when functioning within an organization, can afford to be too much led by personal preferences. The manager in our case would use financial, strategic or operational arguments to justify his decision in order to, for example, convince the bank to lend him money for the purchase. Thus, an importance assessment calls upon the ability to solve logical problems, yet it always includes dealing with affective elements as well. The amount of expertise the assessor has does not matter in this respect. Both novices as well as experts are subject to this personal element in the decision making process.

We are especially interested in decisions where one can expect that the degree of expertise influences the way in which the importance assessment is made. On the other hand, the decisions concerned should not be routine decisions, whether the actor is a layman or an expert. In that situation the decision-maker would be able to fall back on previous importance judgments. Going through the importance assessment process would not be necessary.

This goal sets two important conditions for our study. Firstly, the decision to be studied should be, as stated before, a non-routine decision. We chose the decision described in the introduction: the acquisition of minibuses. Secondly, we had to control the degree of expertise. We did not compare experts with laymen, for reasons explained later. Instead, we studied actors who were not experts in the field of the particular decision situation (in our case: buying minibuses) but who were qualified as decision-makers in the sense that they possessed:

- a general awareness of the field the decision was about;
- general reasoning skills;
- general problem-solving skills;
- experience with self-evaluation (reflection on performance).

Even if actors possess these qualifications, it is of course not certain whether they will use them in a non-routine decision setting and, if so, in what way.

In this article, actors who possess the above-mentioned qualities are called ‘laymen’. We define laymen as actors with little or no experience in the task to be fulfilled. This pertains both to the content of the task, (in our example the task pertained to minibuses) as well as to the procedure to be executed (making an importance assessment). An expert, on the other hand, is a person who has a certain amount of experience with the content of the task (the manager in our example is considered an expert because he is familiar with operating minibuses). The expert may also have experience with the procedure to be executed, but this is not necessary, and in this article we focus on actors for whom the task is not routine. So, whatever experience they have with making importance assessments in the case of minibuses should be limited. The manager in our example may know a lot about operating minibuses, but he has very limited experience, perhaps none at all, in assessing the importance of safety versus comfort. This definition corresponds with the characteristics of experts common in the literature [2], with the exception that most literature we find no combination of a high level of content expertise with a non-routine problem. In other words, in most studies of expertise, experts are assumed to have both the content (domain) knowledge and are familiar with the problem at hand.

In the next section we will start with a brief theoretical overview on the process of importance assessments and judgments (for an elaborated version see the PhD thesis by Heerkens, 2003). Subsequently, we will go into the theory on expertise. Expectations are then formulated with regards to certain aspects of expertise (Section 3), followed by the research methodology (Section 4) and the resulting model of the weight assessment model (Section 5). Expectations are tested in Section 6. Finally, we will go into some limitations of the study and we will suggest an agenda for further research (Section 7).

2: Theoretical framework

Scientific relevance

Our study is new in several aspects.

Existing literature [3] provides a general indication of phases in problem solving. However, no specific model for the phases in importance assessment exists. Various normative models for eliciting weights (importance judgments are available [4], as is a vast body of knowledge on the behaviour of experts (for an overview, see [1]). But the literature concerns the result of importance assessment (the weights elicited from and used by decision-makers) and behaviour of experts solving logical, analytical problems. The way importance assessments are made remains a ‘black box’. Understanding the importance assessment process is vital if we want to give decision-makers instruments to help them with these assessments.

Most research on expertise has concentrated on cognitive issues like the mental representation of problems, the role of domain-specific knowledge, the use of solution algorithms and heuristics and the role of logical reasoning [2, 5, 6, 7]. Insofar as importance assessments have been addressed, they were either seen as ‘black boxes’ or analysed in the context of logical reasoning or external factors. Examples of research questions that have been addressed are: ‘Which arguments do experts and laymen have for their importance judgments?’ and ‘To what extent does domain knowledge influence the judgment?’ [8 and 9].

Views on expertise: the linking of expertise and importance assessment

In research on decision-making there are two views about experts. The first view, which emerges from behavioural research on decision-making, is sceptic about experts. Data suggest that experts from a wide range of expertise domains are not much better predictors than less experienced people. Furthermore, this view suggests a simple technology for replacing experts, i.e. a linear regression model (for example in the domain of medical diagnosis, using illness symptoms and medical judgments as inputs) [10]. The regression does not mimic the thought process of an expert, but it probably makes *more* accurate predictions than the expert does. In some cases a linear regression model containing only three or four variables may accurately predict experts’ behaviour [11] despite experts often believing that they take into account a large number of factors in their decisions.

The second view, stemming from research in cognitive science, suggests that expertise is a rare skill that is developed only after much instruction, practice, and experience. The cognition of experts is more sophisticated than that of laymen; this sophistication is presumed to produce better predictions. This view suggests a model that strives to mimic the decision policies of experts, i.e. an 'expert (or knowledge-based) system containing lists of rules experts use. Such an expert system tries to match, not exceed, the performance of the expert it represents.

Whereas behavioural decision theory emphasized the *performance* of experts, cognitive science usually emphasizes differences in experts' *processes* [8]. In this contribution we use the cognitive approach. The main justification for our approach lies in the fact that expertise and expert behaviour can only be understood by studying the phenomena as they are, in their full complexity, and not by 'fragmentations' into 'simplified' research questions. Observing the behaviour by means of think-aloud protocols enables us to uninhibitedly take stock of different (chains of) phases respondents go through while making an importance assessment. Also, we can analyse the approach of each subject in terms of expertise-related characteristics. Thus in our approach process and expertise characteristics are related. As far as we know, identification of these relationships has not been achieved before.

Research on dimensions of expertise

Much research has been done on the way experts such as doctors, investment brokers, judges, typists and pilots perform their tasks, often focusing on differences between laymen and experts (see for example [2, 5, 6, 7 and 12]). Lots of efforts have been made to develop a general theory of expertise. Sternberg [13] has tried to place various European and American contributions on expertise in a more general framework that encompasses all of the various conceptions or aspects of expertise. Glaser [14] was the first who tried to summarize his thoughts on expertise in a set of propositions. He succeeded in encapsulating his thoughts in a set of more or less general statements. Expertise is the result of a continuous development process. It is very specific and development of expertise is influenced by task demands in the course of experience, to mention but a few.

Although the picture of expertise is biased on account of the highly structured domains in which it has been studied, compared with the study of expertise in complex professional domains like management wherein periodically importance judgements are asked for, it has proved to be very useful as a starting-point for the understanding of the phenomenon and to discover its compiling ingredients [1, 12].

Theoretical and empirical controversy abounds with regard to the understanding and potentials of the concept of expertise. The divergent meanings attached to the concept of expertise create great confusion mainly owing to the domain-specific character of expert behaviour (c.f. [15, 16 and 17]). Because of the fact that most of the researchers restrict themselves to one domain of expertise and try to examine outstanding behaviour in that particular domain [2, 6 and 7], the understanding of the concept in general is greatly in need of elaboration. A compilation of expertise research and subsistent ingredients that altogether compose professional expertise was made by Van der Heijden [1, 12]. Subsequently, the concept of professional expertise has been operationalized in a multi-dimensional way.

Results of studies on expertise seem to show unanimously and quite clearly that expertise comprises

- 1: a *knowledge* component,
- 2: a *skills* component and
- 3: a *meta-cognitive knowledge* component.

In these components, the qualification requirements for decision-makers can be recognized. General field awareness falls under the knowledge component. Reasoning and problem-solving skills are part of the skills component. Experience with self-evaluation falls under the meta-cognitive component.

Nonetheless, the study by Van der Heijden [1 and 12] among professionals from different job fields, made it clear that these three components are not sufficient to cover the phenomenon of professional expertise in its entirety. Measurement of cognitive abilities and skills is not enough to

fully cover the construct (see also [18 and 19]). Motivational aspects and self-insight, as well as social skills, social recognition and growth capacities are important interactors and moderators. That is to say, there is a compelling reason for the proposal of a broader type of measurement, in which cognitive abilities and overt skills play an explicit, but partial role [1, 12]. For sake of clarification the different dimensions will be explained briefly.

The *knowledge* dimension comprises the different types of knowledge that are inherent to a certain professional field. These different types of knowledge are declarative knowledge ('knowing that'), procedural knowledge ('knowing how') and conditional knowledge ('knowing when and where or under what conditions') [20].

The second (skills) dimension has to do with the skills an employee needs to perform the required professional tasks. Once the activities and responsibilities have been defined, it is clear which skills are necessary to perform a given job. A person can only be referred to as an expert if his or her overt behaviour demonstrates the capacity to perform qualitatively well in a particular domain in terms of benchmarking with colleagues. This component of professional expertise is called the dimension of professional *skills*. These skills can pertain to the content of the area of expertise, like being able to work with certain software, and process skills, like communication (writing) and project management skills.

This third dimension is the so-called *meta-cognitive knowledge* dimension ('knowing about knowing' or 'knowing that one knows'). This dimension, that has to do with self-insight or self-consciousness, is known by a wide variety of names: meta-knowledge, executive control knowledge, self-knowledge, regulative knowledge and meta-cognitive strategic knowledge, to mention but a few.

Yet, notwithstanding the importance of knowledge and skills professional expertise can only exist by virtue of being respected by knowledgeable people in the work setting. Even where people are to a large extent endowed with different types of knowledge, with capability for self-regulation and with the capacity to perform quickly and competently, they may not be labelled as specialists. It seems reasonable to presume that other mechanisms play a part in the assessment of people as to whether they are talented or not. It is very important to be able to develop impression management skills, social intelligence and communication skills. That is, the fourth dimension of expertise: *acquisition of social recognition* is of high importance in professional settings. As will be shown, in this study this dimension boils down to *presentation skills*.

A fifth dimension that has been identified by Van der Heijden is the dimension of *growth and flexibility*. People who are capable of acquiring more than one area of expertise within adjacent or radically different fields, or who are capable of acquiring a strategy to master a new area of expertise or expert performance in another territory can be termed 'flexexperts' [21]. These are people who are both flexible and in possession of expertise. In terms of decision theory, growth can be labelled as *learning*.

The central characteristics of professional competence or expertise in current working life are change-orientation and multi-dimensionality [22]. The adaptability and flexibility requirements of an individual employee are based on the continuous changes and transformations taking place at different levels (the individual worker, the job organization and the labour market policy) and on their interaction as well as on global developments, such as that of technology and economy.

To prevent unnecessary elaborations, a short outline has been given here. For the interested reader, we refer to an elaborated outline in the thesis by Van der Heijden [1] and in the article on the psychometric studies that indicate satisfying reliability and validity of the five-scale instrument [12]. In this contribution we aim to judge the benefit of activities in the assessment process by placing them into the multi-dimensional expertise framework by Van der Heijden, with emphasis on the aspects 1, 2, 3 and 4.

3: Expectations concerning the weight assessment process from the perspective of expertise

Based on the theoretical insights discussed above, some expectations are formulated concerning importance assessment in the context of non-routine problems. We speak of expectations and not of theories because we were not able to assure statistical validity, due to the small number of subjects.

This study concerns individual weight assessment processes at a single point in time, for reasons discussed later. So the first three of Van der Heijden's dimensions were the most important.

No expectations were formulated concerning social recognition in general. But actors may be expected to consider how the results of their work are received in an organizational environment, even if we do not study the functioning of the actor within organizations. This means that the dimension of social recognition boils down to the preparation of the *presentation* of the results in a way that maximizes recognition of the quality of the importance assessment performed by an actor. The fifth dimension, learning, was not an issue in our study. Apart from the unintentional learning by fulfilling the assignment, we assume learning did not take place.

Before we address expectations concerning Van der Heijden's dimensions of expertise, which pertain to specific elements of the importance assessment process, we formulate an expectation about the structure of the importance assessment process as a whole.

The general structure of the assessment process

As stated earlier, our study concerns laymen who are qualified as decision makers because they possess, amongst other things, general problem-solving skills. It seems logical that actors possessing these skills will use them to devise a solution method for a problem that they have no previous experience with. After all, these skills are all they have readily available to solve the problem. When operating in an organizational context, not explicitly or implicitly devising some sort of structured way to approach the problem does not seem to be an option, for the results are likely to have to be communicated to - and judged by - other actors in the organization. So, problem-solving skills can play an important part in structuring importance assessments. We expect to be able to detect a use of problem-solving skills by identifying a general layout - to be formalized in a model - of the way actors perform importance assessments. Once we succeed in constructing a model of the importance assessment process, we expect phases can be recognized analogous to those found in problem-solving models. In the literature, the problem-solving approach is an established way of describing human mental activities [3, 13, 23, 24, 25, 26]. It will become clear that the model we have developed in order to test expectation 0 will allow us to set the analytical framework for dealing with subsequent expectations.

Therefore, we formulate the following expectation:

Expectation 0: Subjects use general problem-solving strategies.

Indicators for this expectation are:

- the use of a problem-solving approach in which the phases of a general problem-solving model can be recognized. It will become clear that our model of the importance assessment process largely follows these phases.
- the decomposition of the problem in sub-problems. An obvious manifestation would be the decomposition of attributes in sub-attributes.

Knowledge

Van der Heijden [1] describes as indicators of the knowledge dimension relevant for our study: the ability to apply knowledge to new, unfamiliar problems, having technical knowledge (what we earlier called field awareness) and having an arsenal of problem-solving strategies. The actors we study score low on these indicators. Their reasoning skills may help them solving new problems, but they have only rudimentary field awareness. There is no reason to suppose that actors facing non-routine problems have an arsenal of problem-solving strategies available for it. This applies especially to making explicit and motivated importance assessment for an organization. The actors may have a feeling about the importance of attributes, and may be able to generate arguments so support their judgments, but they don't have a dedicated structure to fit their cognitive activities in.

This lack of a dedicated structure means that the actors are likely not to have a clear idea about the attributes to be weighted. There are several ways the actors can clarify the nature and meaning of the attributes: by formally defining them or by splitting, or decomposing, them into sub-attributes (as already mentioned above). We had no idea how whether actors would use definitions, but there is

literature available on the decomposing of attributes in importance assessment tasks [27]. Because actors are not expected to have a clear idea of the meaning of the attributes to be weighted, we expect them, if they decomposed attributes, to generate large numbers, and a great variety, of sub-attributes, so as to avoid having to attach specific, narrow meanings to attributes. Now what is 'large'? Research indicates that decisions of actors can to a great extent be predicted accurately using a linear additive model containing only three to four attributes [11], although actors generally think that they take many attributes into consideration. By analogy, one would expect that three to four sub-attributes would be sufficient to capture the meaning of an attribute accurately enough to serve as a basis for decisions. If we assume actors use more sub-attributes than necessary, then we feel that in our test we can settle for a three- to fourfold redundancy in sub-attributes as an operationalization of "more than necessary", so we take the round number of ten sub-attributes per attribute.

Therefore, we formulated the following expectations.

Expectation 1: Subjects take into account at least ten sub-attributes for each of the two attributes to be weighted

The indicator for testing this expectation is the number of aspects (sub-attributes) that safety and comfort are decomposed into.

Facing non-routine problems, the actors that are the subjects of our study lack the insight in the problem area needed to analyse causal relationships between (sub-)attributes. After all, they have only general field awareness. Establishing causal relationships can be relevant for weighting. It can lead to the use of a common denominator, like money. The relative effects of safety and comfort on a variable related to money, like revenues of the minibuss company, could be used as an indicator of the attributes' importance. Establishing causal relations can also be used for eliminating redundant attributes. For example; if braking distance is a sub-attribute of safety and weight determines braking distance, it is useless, even dysfunctional, to consider both weight and braking distance in assessing the safety of a minibuss. Finally, insight in causal relationships can be used as a form of framing the problem, getting a 'mental model' of what 'safety' and 'comfort' actually mean.

Expectation 2: Subjects will not concern themselves with explicit causal relationships between (sub-)attributes.

Several indicators of the concern for causal relationships are possible, like the number of segments in the think-aloud protocols pertaining to causal relationships and the explicit use of methods for establishing such relationships, like cognitive mapping. We don't use the first indicator because it says nothing about the number and importance of causal relationships eventually established, but we use the second one. We also use the extent of integration (taking together) of sub-attributes. If sub-attributes A and B are taken together as attribute C, we assume that subjects establish a causal relationship between A and C and B and C, respectively. In the example of the minibuss decision: if 'quality of brakes' and 'acceleration' are integrated as 'active safety', then we assume that the quality of the brakes and the acceleration that the minibuss can attain determine active safety.

As stated, for expectation 1, we used as an indicator the number of sub-attributes in which an attribute is decomposed. Whereas there is an obvious relationship between an attribute and its sub-attributes, this need not be a causal relationship but may merely be an implication that there is some sort of a relationship, or a specification relationship (the sub-attribute is a specific form of the main attribute). So, we don't use the number of decompositions as an indicator for expectation 2.

Actors possessing only rudimentary field awareness would not have sufficient knowledge (let alone skills) for applying methods and techniques specifically aimed at making importance assessments. Such methods and techniques include: the many methods developed to elicit weights that people attach to attributes (sometimes without themselves knowing it). It would, for example, be possible for actors to imagine for themselves choices between a number of alternatives with varying

values on a number of attributes, and then, based on these choices, infer the weights that they use. We do not expect actors dealing with non-routine problems to use such sophisticated methods.

Expectation 3: Subjects do not use methods and techniques specifically aimed at making importance assessments.

As indicators we used firstly the explicit mentioning and subsequent use of specific methods, or the use without explicit mentioning, but in that case the use had to have some level of consistency. For example, some weight elicitation methods involve pairwise comparison of (sub-)attributes. Now if expectation 1 can be accepted and subjects indeed generate many sub-attributes, there are bound to be some incidental pairwise comparisons. But to qualify as a method, there should be consistent pairwise comparison, with a clear conclusion in the end concerning the weight values.

Utility theory states that the attractiveness of an alternative is expressed in the summation of scores on the various attributes, multiplied by their weights [4]. The product of the score and weight of an attribute is called the partial utility of that attribute. The sum of the partial utilities is the total utility (attractiveness) of the alternative. So, utility theory assumes that a person judging an alternative has some concept of (partial) utility that makes various attributes comparable to each other. When utility theory is applied properly, the scores of all attributes are expressed in the same dimension, or on a common dimensionless scale, like a five-point scale. So, all possible scores of each attribute are translated to a score of, for example, 1 to 5. This would mean that weights are also expressed on a common scale. After all, the products of scores and weights all have the same dimension: (partial) utility. In that case, the importance of attributes need not be established by comparing them and deciding whether one attribute is more important than the other. One could just attach an importance value to each attribute individually, just like you can see from the speedometer how fast a car is going without having to compare its speed to that of other cars on the road.

For actors dealing with non-routine problems, absolute weighting has the advantage that no difficult direct comparisons with other attributes are required. The comparison of attribute importance follows from the weights themselves. For example, if safety gets absolute weight 0.4 and comfort 0.6, then comfort is more important than safety even if they have not been directly compared. Of course, comparison could lead to adjustment of the weights, but that is optional. Also, questions could be raised about the validity of weights given without comparison to other attributes, but that need not make the method less attractive for actors facing importance assessment problems that may be difficult for them to solve anyway.

So, if we assume that actors reason according utility theory, we could formulate the following expectation:

Expectation 4: Subjects express the importance of each attribute on an absolute scale.

As an indicator we used the average amount of effort subjects spent on weighting attributes in isolation ('safety is important') versus comparing the importance of attributes ('safety is more important than comfort'). But this is not enough. Actors can weigh attributes in isolation and yet use different scales for each attribute (whether this is sensible or not). They may even use identical or similar scale values like 'quite important', 'really important' and the like without meaning the same in each case. So, as a second indicator we look at the presence of a common scale, either by a consistent use of scale values (a certain value indeed consistently means the same level of importance) or by the presence of an explicit algorithm according to which absolute weights are converted into relative ones. Such an algorithm is not possible if there isn't some sort of common scale, or scales convertible into each other.

Skills

Van der Heijden [4] formulated several indicators of skills that are appropriate for this study. They are: the ability to use specific methods and techniques and to work independently, and the general skill level as perceived by the actor.

One type of skills: general problem-solving skills, has already been covered in expectation 4. If this expectation is accepted, another expectation is logical. If actors adopt a phased problem-solving approach, it would be interesting to study the amount of effort spent on structuring the problem versus the effort spent on solving it. It would be reasonable to assume that actors having limited field awareness and problem-solving skills confronted with a non-routine problem would spend relatively much effort on structuring the problem. Kotovsky & Fallside [27] indeed provide proof for this. It is difficult to quantify the effort spent on the two activities. Kotovsky & Fallside studied a totally different type of problem (strictly logical problems) and provide no usable quantitative clues. We settled for a simple 50% rule: actors spend at least as much effort on structuring as on solving the problem.

Therefore we expected the following:

Expectation 5: Subjects spend at least as much effort on structuring the non-routine assessment problem as they do on solving the problem.

The indicator is the average percentages of statements pertaining to the structuring phases of our model compared to the solving phases.

Not only at the beginning of the importance assessment process, but also at the end, the characteristics of the actors we studied lead to an expectation. We can expect that actors are well aware of their lack of proficiency in solving a non-routine assessment problem. Therefore, it seems reasonable that they spend quite a lot of effort on evaluating their work. They will likely not readily accept their way of working, or the result of their work, as adequate, but they will try to assess whether they actually got it right. We expect them to spend a significant amount of effort on evaluation. As to what is 'significant', we decided on a threshold of the average effort per phase, the exact number to be decided upon after the model of the importance assessment had been constructed.

Expectation 5': Subjects spend an above-average amount of effort on evaluating the quality of their work.

The indicator we used is the average number of segments in the think-aloud protocols pertaining to evaluation (as we will see, evaluation was identified as a phase of importance assessment).

Meta-cognition and presentation

Van der Heijden's [1] indicators of metacognition relevant for this study are: capability of judging the level of one's own knowledge and skills in a particular problem situation, capability to judge one's own ability of separating main issues from details, to weigh pros and cons of working methods and techniques, to keep oversight over complex situations and the ability to judge which skills one is missing when confronted with a new problem. We did not measure these indicators but assessed the confidence the subjects had in their own importance judgment. The actors we studied have, as we have seen, little, if any, experience with the task given to them, but they possessed self-evaluation experience. So, the actors can be expected to have adequate cognitive abilities to assess their own performance. This leads us to expectation 6:

Expectation 6: The subjects have little confidence in their own judgments (since they are very aware of their lack of knowledge and skills to solve the non-routine problem presented to them).

But an alternative line of reasoning is possible. Actors involved in acquisition decisions may possess a considerable amount of training in project management, communication and presentation of their work. These characteristics were not assumed beforehand, but they may be there nonetheless. It is possible that they, while realizing their lack of content knowledge and specific importance assessment

skills, they are quite confident in their project management skills required to handle the assignment properly, and in their presentation skills, needed to convince others of the quality of their work. This, combined with the not unreasonable notion that importance judgments are to a certain extent a matter of personal opinion anyway, gives rise to the assumption they may well have considerable confidence in the quality of their work. We enter the realm of van der Heijden's fourth dimension: social recognition, which we have narrowed to presentation.

So, an alternative to expectation 6 is:

Expectation 6': Subjects are confident in their weight judgements (based on the project management communication and presentation skills they possess).

The indicators for these two expectations are the confidence in their judgments subjects expressed in the interview after having concluded the assignment, the frequency of preference reversal, the acceptance of expectation 4 as an indicator of the use of project management skills and reflections by the subjects on their performance made during the interview afterwards.

Each of our expectations will be investigated in Section 6.

4: Research methodology

Only a short outline of our research methodology is given here. For an extensive description we refer to Heerkens (2003).

Sample and assignment

18 undergraduate students of the University of Twente were given an individual assignment based on a fictional case similar to that used in the introduction to this article. University students might be assumed to have enough analytical abilities to perform the assignment satisfactorily. At the university, all students receive basic training in general problem-solving techniques. They are specifically indoctrinated in the importance of having a proper problem definition (splitting the problem if necessary), in explicitly choosing a problem-solving method and evaluating the solution. Yet, they don't have enough knowledge and skills in this area that would enable them to rely on previous experience of importance assessments. Hence, the danger that they give weights based on previously obtained knowledge is minimized. This is what we wanted, given that our study is about non-routine decisions. The students had extensive experience with projects concerning the solving of management problems, and insight in the theory of management. So they can be assumed to have some feeling for operating within a simulated organizational context.

Their assignment consisted of supporting the acquisition process of new minibuses by a local company. The subjects were asked to establish the importance of two characteristics of the to-be-acquired minibuses *vis-à-vis* each other, and were asked to imagine that they would be advising the management team during the acquisition process. The attributes, *safety* and *passenger comfort*, were chosen to prevent comparability by some readily available algorithm or heuristic or easy expression in a common denominator such as money. The information that was supplied included a brochure of the company, a leaflet explaining the decision context and two brochures on minibuses; one on a Volkswagen and one on an Opel. The latter enabled the subjects to get familiar with the specific capital good to be acquired. It was made clear that these examples of minibuses did not mean that the subjects had to make a choice between them. A pre-set structure of (sub-)attributes was avoided as much as possible.

The minibus problem was chosen because subjects were assumed to have superficial familiarity with them, so that they could give an importance judgment at all. With, for example, airplanes this would likely have been much more difficult.

Procedure

The respondents were asked to think aloud during the assessment process. The general guidelines for think-aloud studies given by Ericsson & Simon [25] were followed, including a practice session to familiarize the subjects with the think-aloud strategy. All verbal information given by the respondent was recorded and typed out verbatim. After completion of the assignment, a short interview was conducted. In total, each session lasted for a maximum of two hours, for which the subjects were paid 20 Euro. Two pilot sessions were conducted, which led to some minor adjustments of the assignment. Two kinds of analyses have been performed using the typed out protocols:

1: A largely qualitative analysis according to the general rules of the ‘Grounded Theory’ approach [28]. This was aimed at discovering regularities in the protocols by inferring general working rules from the think-aloud protocols and in some cases determining the percentage of subjects applying certain rules.

Using this analysis, several phases in the importance assessment process could be distinguished. They will be discussed in the next section.

2: A more quantitative analysis based on a formal coding scheme that was designed on the basis of the qualitative analysis. As mentioned above, certain phases were discovered during the qualitative analysis. In the quantitative analysis the effort put into each phase was assessed, expressed in the number of verbal protocol segments devoted to each phase. Two coders performed the coding activities. Although they worked independently of each other, during the coding of the first six protocols weekly meetings were held to discuss general coding issues in order to enhance its reliability. The coders retrospectively applied the refinements to the coding scheme independently. The overall Cohen’s Kappa [29] for inter-rater consistence was .97 over a total number of verbal segments of 7253.

The expectations are largely analysed in quantitative terms. Mostly, we used average numbers across all subjects in testing the expectations. For expectations pertaining to individual behaviour as observed, we use the following acceptance criterion. We accept an expectation if 80% of the subjects (15, rounded off upwards) behaves according to the expectation. For negatively formulated expectations, we reject them if 15 subjects act contrary to the expectation. This is a somewhat arbitrary threshold, but with the small number of subjects involved, this seems a reasonable level of significance. Sometimes some additional criteria are taken into account; where appropriate this will be mentioned. In cases where it was not a case of an indicator either present or not present with individual subjects, but where the value of an indicator could be measured quantitatively, average values were used as criteria.

In the next sections we will go into the results of the think-aloud study.

5: Results: the Weight Assessment Model (WAM)

Firstly, the phase model that has been derived from the data is given. In the next section *the way attributes are processed and the confidence in the end result* will be discussed.

Our Weight Assessment Model (WAM) consists of 7 main phases and 6 auxiliary activities. The 7 main phases are presented in a sequential way in table 1. In reality, subjects often go back and forth between phases and most often address phases more than once. The auxiliary activities pertain to areas like information search and planning. We do not deal with them in the context of this article (see Heerkens, 2003).

The WAM follows the general format of phased problem-solving models [3, 13, 23, 24, 25, 26, 30]: problem structuring, often called problem identification (phase 1), problem analysis (phase 2), problem solution (phases 3 to 6) and evaluation (phase 7). In many models an implementation phase is also found, but as our problem had no implementation component this phase is not relevant in this study.

From the table, it is clear that most phases could be observed with every subject. Exceptions were phases 4 and 5. In each phase subjects could be found who had either 5% or more of their total number of statements, including auxiliary activities, devoted to that phase, or at least 30 statements.

Thus we conclude that it is justified to distinguish these phases and that they are not just isolated coding incidents.

Let us now discuss briefly the different phases of the model. Only the details relevant for testing the expectations are covered here. For a more detailed explanation, we refer to Heerkens (2003).

Phases of the WAM

Phase 1: Problem identification

This phase consists of activities like reading the assignment and, if desired, re-formulating it in one's own words. When re-formulating aspects of the assignment, subjects may also mention things they are *not* to do, for example weight attributes concerned with costs. Essentially, this phase concerns defining, so to speak, the task lying ahead.

Phase 2: (Sub-) attribute processing

We find aspects of problem analysis in phase 2. Not only the elements constituting the problem are defined (elaboration on the character of the attributes to be weighted), but also the possible, not necessarily causal, relationships between these elements.

If one wants to weight attributes, one should first know what one is weighting. Attribute processing concerns giving the attributes a more precise meaning. This can be seen as a case of framing [31].

Attribute properties like measuring level, measuring unit, level of abstractness and precision can change as a result of processing. The following forms of processing were identified:

a: Decomposing. An attribute can be split up in several sub-attributes. For example: 'safety' can be split up in 'braking distance', 'acceleration', 'strength of the structure', 'presence of seatbelts' and so forth. Each of these sub-attributes can, of course, be decomposed further.

b: Re-formulating. When an actor gives an attribute or sub-attribute a different name while meaning the same attribute with a similar, not necessarily identical, measurement unit, the attribute is re-formulated. For example, 'comfort' may be re-formulated as 'passenger comfort', 'comfort of the passengers' or 'travelling comfort'.

c: Specifying an attribute or sub-attribute. For example 'the minibus should be roomy enough' might be made concrete, as 'people should not have to sit with their suitcases on their laps'.

d: Integrating (sub-)attributes into a new 'sub-attribute'. For example: after having named sub-attributes of safety like 'braking distance', 'acceleration', 'strength of the structure' and 'presence of seatbelts', the first two could be taken together (integrated) into 'active safety' and the last two could be integrated into 'passive safety'.

e: Making an attribute more abstract. This is the complement of specification.

The next phases concern the actual weighting process.

Phase 3: Absolute (sub)-attribute weighting

With 'absolute' weighting [32], we mean that a statement about the importance of a (sub)-attribute is made without reference to the importance of other (sub)-attributes. For example, if an actor says: "safety is important", or "having seatbelts in the minibus is absolutely essential", it is not clear whether safety is more important than comfort, or whether having seatbelts is more essential than having an anti-skid system.

In phase three, as well as in the subsequent phases, subjects not only assign weights, but they also give the reason for giving these weights. For example, a subject may say: "I think a good anti-skid system is important (weight assignment), because (argument) it reduces the chances of getting involved in an accident". It is possible to decide that some sub-attributes defined earlier, are to be excluded from the weighting activities in subsequent phases. So, this phase, like phase two, can serve to give the subject a clearer view of the (sub)-attributes to be weighted.

Phase 4: Homogeneous sub-attribute weighting

This phase is the first in which 'true' weighting takes place; i.e. the balancing of the weight of one sub-attribute against that of another. We call this 'relative weighting'. In this phase, two or more sub-attributes of the same main attribute are weighted against each other, and arguments for the weighting

are given. For example, the presence of seatbelts, the quality of an anti-skid system and the strength of the structure (all sub-attributes of safety) may be weighted against each other.

Phase 5: Heterogeneous sub-attribute weighting

This phase differs in only one respect from the previous one, i.e. the sub-attributes that are weighted belong to different main attributes. For example, the quality of the anti-skid system (a sub-attribute of safety) might be weighted against the quality of the seats (a sub-attribute of comfort). The weight qualifications are the same as in phase 4.

Phase 6: Attribute weighting

This phase concerned the integral weighting of the (two) main attributes. The weight qualifications appeared to be the same as in previous phases. This was the original assignment. Only 13 of the 18 subjects reached this phase. 5 stopped after having completed phase 5. Some of these subjects consistently judged sub-attributes of safety to be more important than sub-attributes of comfort, or the other way around. In that case, a judgment of the main attributes can be readily inferred. But this does not lead to the conclusion that subjects actually completed phase 6, for they may not have made the inference we made. Some of the subjects who stopped after phase 5 made some statements pertaining to phase 6. For example: at the beginning of the assignment they would remark that, at first sight, safety seemed more important to them than comfort. Hence phase 6 was considered as being addressed (even with 100%), though sometimes with no explicit statements.

Phase 7: Evaluation

This phase comprises the reflections by subjects on their activities and the results. Several types of evaluation were identified, like the extent to which the assignment had been fulfilled, evaluations of weights judgments (was the subject, on hindsight, happy with assigned weights) and evaluations of arguments (how good were the reasons for particular weight judgments).

Use of general problem solving strategies.

Looking at Table 1,

Expectation 0: Subjects use general problem-solving strategies.

can be easily accepted.

All subjects went through the phases of problem structuring, analysis, solution and evaluation, even though some solution phases (in particular homogeneous and heterogeneous sub-attribute weighting) were sometimes rarely used. All subjects split the problem into sub-problem by decomposing the main attributes, as stated above.

Furthermore, phase 2 could clearly be identified as the phase that subjects enter immediately after they have read the assignment (which is part of phase 1). One can easily detect the logic behind this; before one can attach a weight to a (sub)-attribute (phases 3 to 6), one has to identify it and process the different sub-attributes. Usually, phase 3 precedes phases 4 to 6. If subjects assign weights to attributes on an absolute importance scale (expectation 4), then phases 4 and 5 could well be unnecessary, for the subjects could immediately proceed to phase 6. But phases 4 and 5 could be used as checks and refinements on the absolute weights given in phase 3. In that case, the sequence of phase 3, 4 and 5 is also logical. Moreover, the assignment was to compare the importance of safety and comfort. If sub-attributes of safety can be weighted against sub-attributes of comfort (phase 5) then it does not add much to the fulfilment of the assignment to weigh those sub-attributes against other sub-attributes of the same attribute (phase 4). Likewise, if safety and comfort are already weighed (phase 6), the assignment is fulfilled and there is no need for weighting sub-attributes (phases 4 and 5). Phase 6 concerned the desired result of the assignment (weighting safety against comfort), so it is logical that it is the last weighting phase, followed only by evaluation.

This concludes the description of the different phases of the decision making process. In the next section, the other expectations we formulated earlier are tested.

6: The relevance of the WAM in the light of expert behaviour: testing of the expectations

Expectation 1: Subjects take into account at least ten sub-attributes for each of the two attributes to be weighted

This expectation is accepted. As mentioned earlier, attributes could be processed in one of the following ways: decomposing, specifying, re-formulating, abstracting and integrating. The way of processing of attributes has a profound effect on subsequent phases of the weight assessment process. The result of the processing phase can be seen as an intermediate product, i.e. the attributes are formulated in such a way that they can, ideally, be given weights easily, unambiguously and with a clear motivation.

Table 2 shows the number of sub-attributes for 'safety' at various levels of decomposition. A level is defined as the number of splits that resulted in a certain sub-attribute. So, if 'comfort' is split in a number of sub-attributes, amongst which 'quality of the seats, which in turn is split in 'width of the seats' and 'height of the armrests', then there are two levels of decomposition. For comfort, the numbers were roughly similar. It can be seen that the attribute is decomposed in a large number of sub-attributes. The average number of sub-attributes per subject for 'safety' was 19,6. For comfort the number was 24,4. 15 subjects generated 10 or more sub-attributes for safety. All subjects generated 10 or more sub-attributes for comfort. So, we accept expectation 1.

What did the subjects do with all these sub-attributes? One could think that the subjects wanted to use the sub-attributes in order to discover common denominators, for example by identifying sub-attributes that safety and comfort have in common and that hence might be eliminated or given equal weights. Yet, no evidence supports this notion. The decomposition was rather unsystematic. Subjects did not indicate a desire to perform phase 2 in a systematic way neither during the think-aloud session nor during the interview afterwards (although some subjects indicated they would want to do this if they ever had to perform a similar assignment again).

Expectation 2: Subjects will not concern themselves with explicit causal relationships between (sub-) attributes.

This expectation is accepted. No systematic methods for finding causal relationships were used, like cognitive mapping. Not a single subject tried to define or frame the two attributes to be weighted (safety and comfort) in order to express them into a common denominator, for example money or another explicitly mentioned common scale on which the attributes could be scored. Instead, they decomposed the attributes in a large number of sub-attributes, as already discussed.

When a systematic procedure could be inferred for weighting some sub-attributes (for example, the distinction between active and passive safety), this was occasional and not maintained with respect to all relevant sub-attributes. What is more, there was hardly any integration of sub-attributes, something that one expects when weights are to be given not to sub-attributes but to main attributes. This is in line with expectation 2, in the sense that for integration the (causal) relationships between (sub-) attributes need to be known. Indeed, relationships between attributes were addressed only incidentally. If they were discussed, they seldom had consequences for the weighting. Integration was a rare occasion, as can be seen from Table 3. This table shows the number of (sub-) attributes integrated. In the acceptance of expectation 2 we classify a subject as a non-incidental user of integration if integration occurs in at least 4 cases over safety and comfort together. As we see that fewer than 15 subjects integrate at all, expectation 2 is rejected as far as this indicator is concerned.

Another way to look at the significance of integration is to observe how many (sub-) attributes are the result of integration. The maximum number of attributes that were the results of integration

was 4 (1 subject). 2 integrated attributes were found with only 3 subjects. Only in two instances was a sub-attribute resulting from integration given a weight during the final weight assignment. Integration always resulted in a new sub-attribute, not in the main attributes to be weighted according to the assignment (safety and comfort). The logic of the integration was often implicit and nearly always purely qualitative. No indexing or other quantitative methods were used. In sum, integration was by and large irrelevant.

It can be seen that all subjects ended phase 2 with a large number of sub-attributes, with no system to guarantee that the main attributes have been adequately covered, or that assessing the weights of the sub-attributes would in any way be easier than assessing the weights of the main attributes. This is also in line with expectation 2. The relationships between the sub-attributes and main attributes were unclear, so that weights assigned to sub-attributes did not seem to say anything about the weights of the main attributes. It could be assumed that subjects used the processing of attributes mainly for framing purposes, i.e. to find out what 'safety' and 'comfort' actually mean. But since all subjects devoted considerable effort to phase 3 (absolute weighting) the processing of sub-attributes appeared to mean more than just a framing function. The subjects obviously were not only interested in obtaining concepts of 'safety' and 'comfort'. The sub-attributes were important in their own right.

Expectation 3: Subjects do not use methods and techniques specifically aimed at making importance assessments.

Expectation 3 is accepted: more than three subjects violate this negatively formulated expectation. Only a small minority (3 subjects) used some form of pairwise comparison with some consistency. 8 subjects used a method to progressively refine the range of weights taken into consideration, either starting with extreme weights and working towards the middle or starting with equal weights and working towards extremes. This could be classified as a method specifically suitable for weighting. Still the number of subjects using such methods is smaller than the required 15 required to accept that subjects actually do use such methods..

Expectation 4: Subjects express the importance of each attribute on an absolute scale.

Although the results point strongly at the possibility of subjects using absolute scales of importance, they do not allow us to confirm this expectation. Phases 4 and 5 (relative sub-attribute weighting) are insignificant compared to phase 3 (absolute weighting). This gives credence to the notion that phases 4 and 5 are merely used to check the results of phase 3. If phases 4 and 5 would be steps in establishing weights of sub-attributes instead of a mere check, more effort would surely be devoted to these phases. On the other hand, if subjects would use phases 4 and 5 as checks for phase 3, we would expect them to frequently hop between phases 3 and 4 or 3 and 5, probably with evaluation (phase 7) in between. Although the protocols were not analysed in this respect, a cursory look at the graphical representations of phase changes provided no indications that this was the case to any significant extent.

Another observation casting doubt on the use of an absolute scale is that phase 6 takes 12.5% of the effort. While this is less than phase 3 (27%), it represents much more than an incidental activity. The effort per attribute reinforces this notion: only two attributes were weighted in phase 6, while twenty or more attributes were frequently weighted in phase 3. The insignificance of phases 4 and 5 relative to phase 6 may be explained by the fact that the assignment stated quite clearly that safety and comfort had to be weighted. So, weighing sub-attributes was not enough.

Finally, there are no reliable indications that subjects used absolute scales to any extent. 13 of the 18 subjects gave ordinal weights as the end result of the assignment. Even after the experimenters asked the subjects to give quantitative (interval or ratio) weights, 6 subjects persisted in giving ordinal weights. Examples of ordinal weights often used are: 'quite important', 'really important', 'not very important' and the like. It is impossible to say whether these expressions were part of a scale in the sense that their value relative to each other had any consistency. Some subjects used explicit scales, but they were the exception rather than the rule. And even then, these scales were usually used at the

sub-attribute level and only in one case did the subject (when asked during the interview afterwards) link the weights of the sub-attributes directly to the (ordinal) weights of safety and comfort.

The notion of subjects constructing an absolute scale of importance is attractive, but cannot be accepted, nor clearly rejected, by our results. This is an area for further research. Questions present themselves like ‘is there one scale for all attributes or are there several scales, for example one for each set of sub-attributes pertaining to the same main attribute’? The endowment [33, 34] seems to point at the possibility of more than one scale. The price a person is willing to pay for a good is sometimes lower than the sum he wants to receive before giving it up. Both scores and weights could play a role here, but the example shows that the question whether people have an absolute importance scale available is far from answered. But the significance of phase 3 is obvious in our results. The relationship between the phases of the WAM is further discussed in Heerkens (2003).

There is another reason for the comparative rarity of relative weighting that should be considered: the conscious or unconscious desire to reduce complexity. With absolute weighting of n attributes, there are n weights to establish. With weighting of all possible pairs of attributes against each other (the most efficient form of relative weighting) the number of weights becomes $n(n-1)$. So, the weighting becomes much more complex, and even more so when inconsistencies in the weighting have to be addressed (A is more important than B, B is more important than C, C is more important than A). As the complexity of the importance scales do not differ in principle between absolute and relative weighting, absolute weighting yields a reduction in complexity. This explanation cannot be proven, but subjects certainly employed various other means of complexity reduction (Heerkens, 2003).

Expectation 5: Subjects spend at least as much effort on structuring the non-routine assessment problem as they do on solving the problem.

Table 1 shows that, contrary to our expectation, 37.03% of the average effort was devoted to the structuring phases (1 and 2) and 45.79% to the weighting phases. Only 2 of the subjects devoted more time to the structuring phases than to the weighting phases. We have no clear explanation for this. It is possible that experts would devote even less time to the structuring phase. If this were the case, our threshold for accepting expectation 5 may be too high. But for the moment it is rejected.

Expectation 5': Subjects spend an above-average amount of effort on evaluating the quality of their work.

This expectation is confirmed (see Table 1). The model encompassed seven phases, and 15% is one seventh of the segments, rounded off upwards. So, the threshold for accepting the expectation is 15%. The averaged effort devoted to evaluation is more than this. It is impossible to say whether the reason is indeed, as we suspect, the lack of expertise of the subjects in weighting the attributes under consideration, but the notion seems credible.

Confidence in the end result

Expectation 6: The subjects have little confidence in their own judgments (since they are very aware of their lack of knowledge and skills to solve the non-routine problem presented to them).

This expectation is rejected. Although some subjects remarked that assigning weights is to a certain extent an arbitrary process, all subjects showed signs of having confidence in the weights eventually assigned. Some empirical evidence of this is summarized below:

- *Robustness:* After completing the assignment, we asked the subjects why they assigned the particular weights they did and not slightly different weights (for example 0.6 for safety instead of 0.7). All subjects remained with their original weights, although four acknowledged that slightly different weights would also have been an option. Of course it is possible that the subjects wanted to stick to their original weights because they had invested considerable effort into performing the

assignment, but it was pointed out to them explicitly several times that there were no 'right' or 'wrong' answers.

- *Consistent preference evaluation:* Only one subject of the 11 who in the end weighted safety and comfort explicitly as overall concepts (and not only their sub-attributes) showed preference reversal between the first time they made a statement about the relative importance of safety and comfort and the final weighting. So, most subjects who stated at the start of the assignment that safety was more important than comfort would likely stick to this for the remainder of the assignment, although the exact weights might change. This is all the more remarkable because 4 of these subjects changed preferences during the assignment, but in the end came back to their original order of preference.

So, the subjects seemed to have considerable confidence in their own weight assignments. . This is not a sign of lightheartedness as most of them showed expressions indicating that at certain times they experienced difficulty in coping with the assignment. Also, except for one subject, there are no indications that any of the subjects failed to do their utmost to perform the assignment as good as possible. The one exception had such clear-cut ideas about the importance of safety and comfort that he felt he did not have to elaborate on the assignment.

All in all, while the subjects were not experts in the field of minibuses nor did they have experience in performing importance assessments like this, they had confidence in their own work. This is contradictory to expectation 6, based on the expertise framework by Van der Heijden [1,12]. Looking at the second dimension that is distinguished by Van der Heijden having self-confidence with regard to a certain task is an important aspect falling in the category of meta-cognitive knowledge. Having insight into one's strengths and weaknesses enables a person to evaluate his or her own decision process. And our subjects could rely only to a limited extent to content knowledge (a laymen's knowledge on minibuses) and even less on meta-cognitive knowledge.

So, these subjects may not have been capable to assess the validity of their own judgements, i.e. in this case they are subject to the leniency effect [35]. They are inclined to give a more rosy image, i.e. more faithful judgement of their own decision making process. Besides, our subjects did not need to be concerned with the approval of bosses or colleagues, so their self-confidence was not hindered by fear for a lack of social recognition. In real-life professional settings, which the model by Van der Heijden relates to, gaining respect and approval by others is important. So, the confidence of the subjects might very likely have been lower in real professional settings.

The above explanation of the results is based on the negative relationship between the capabilities for metacognition and confidence assumed in expectation 6. But it is possible that confidence in the end result is high despite a high level of metacognition. Such an alternative explanation is given in expectation Z:

Expectation Z: Subjects are confident in their weight judgements (based on the project management, communication and presentation skills they possess).

This expectation is confirmed in the sense that the indicators discussed above were present. However, alternative explanations cannot be ruled out. We already saw that a general problem-solving strategy could be recognized in the subjects' work. In another article (Heerkens, 2003) we convincingly argue that complexity-reduction strategies (which can be seen as a type of project management strategies) were employed, but in the interviews afterwards subjects did not indicate that these strategies had increased or assured their confidence in the end result. They could have felt this way, but it cannot be proven. There were indications that subjects realized the communicative aspects of the assignments: 3 stated that their own personal opinion didn't matter but that they had to look at the problem from the perspective of the company. All subjects took perspectives of interested parties other than themselves into account (clients, the management of the company, the drivers of the minibuses). Some of them indicated explicitly that they felt that they could justify their importance judgment to the management of the minibus company. But with respect to this expectation, the relationships with project management, communication and presentation skills remain somewhat speculative.

7: Conclusions and recommendations for further research

Based on Simons [3] general problem-solving model, we developed a phase model for representing the way individual actors assess the importance of attributes of capital goods in non-routine situations. The model is consistent with the problem space for weight assessment problems devised on the basis of Simons model and on the elements of a generic linear additive utility function. In this article, we addressed some salient observations done during the think-aloud study.

The results of our think-aloud study show, not surprisingly, that apart from the WAM phase structure, the subjects did not assess weights in a systematic way. They did not in any explicit way try to relate the various attributes to each other, either by developing common scales to score the attributes or by seeking a common denominator, like money. They hardly integrated the many sub-attributes they generated. Neither did subjects use a system in decomposing 'safety' and 'comfort' to uncover incompleteness and redundancies that also might have served as a basis for subsequent integration. Furthermore, they paid very little attention to weighting the sub-attributes against each other (phases 4 and 5 of the Weight Assessment Model), and did not pay too much attention to phase 6 (overall attribute weighting, accept expectation 5'). They did explore the problem broadly (expectation 1 accepted), did not concern themselves with causal relationships (expectation 2 accepted) and used elements of general but not clearly of specific problem-solving methods (expectations 0 and 3 accepted). They emphasized problem structuring less than expected (expectation 5 rejected). Altogether, this leaves us with the impression that non-routine decision problems are tackled using an intuitive approach. All the same, they had confidence in the end result (expectation 6 rejected) but this could not be attributed unambiguously to their project management, communication and presentation skills (even though expectation 6' is confirmed). Considerable effort was devoted to evaluation (expectation 5' accepted).

Our results are quite unambiguous in terms of internal validity. The value of Cohen's Kappa is high. Furthermore, the assignment was, while difficult, not beyond the subjects' ability and all subjects seem to have done their best to perform the assignment as well as they could. The reason for this assessment is that no subject took the easy way out in case of lack of confidence by stating that safety and comfort were equally important. One subject arrived at equal weights but he had a good reason; he combined the weights that he would assign personally with the weight he thought that clients would give.

However, a further discussion of the external validity and the scope of generalization is necessary. The question is to which degree the setting of the study and the subjects were truly representative of real life decision situations.

External validity

Real-life decision processes are much more complex than those studied in our study. For example, more than one actor is likely to be involved. Furthermore, it is to be expected that actors will make weight assessments not only before but also during the choice of alternatives. Also, it is likely that there are more than two attributes, and contextual aspects like the preferences of the management have to be taken into account by the actors involved. A major shortcoming in many expertise studies is the fact that it is debatable whether the standardized laboratory tasks that are mainly used in expertise research capture the real-life problems that people encounter in work settings. This means that there are some validity problems inherent to the choice of the performance tasks that have been studied. Although we have tried to counter the problem of previous experience, which would give some subjects an advantage compared with others by taking non-experienced students who did not have the expert's domain-specific knowledge, the situation lacks 'real-life character'.

In order to make a start on addressing these problems, a study should be attempted of the phenomenon of professional expertise in its entirety and in different real-life settings. One could consider studying the process of importance judgements among more or less experienced managers in working organizations.

Another limitation is that students were used as subjects. They had no experience in making weight assessments for the acquisition of capital goods, except perhaps goods for their own personal use. In organizations, major decisions are often taken by groups of actors who communicate their

importance judgments and their preferences for certain alternatives to each other, and who hopefully in the end arrive at a decision that everyone can live with. But actors bring their individual judgments and preferences to the table, perhaps having entertained them long before the start of the formal decision process. It is on these individual actors that we focus. This is partly a matter of prudence. Our priority is studying the reasoning processes of actors. If we would study groups of actors, it would be practically impossible to observe individual actors closely enough to assess their reasoning processes. Furthermore, it would be hard to distinguish individual reasoning processes from group dynamic processes. Our research method is well suited for studying reasoning processes of individual actors in detail, but unsuited for studying group processes. Once a model for representing individual reasoning processes is available, group decision processes may be studied using it.

The last limitation is that we do not deal with limitations that the organizational context may impose on the (perceived) freedom of an actor in his or her importance assessment process. For example, the opinion of one's superiors may have a profound influence on one's importance judgment. In this study, the actors face no pre-set limits on their freedom of reasoning. This may be an issue for future research.

The limitations that are mentioned previously were deliberately accepted and can also be turned into strengths. What we have tried to do is to establish a baseline case: how do actors make weight assessments when they are not influenced by knowledge about the available alternatives, by an organizational context or by past weight assessments which might serve as a basis for undue rationalization? The students that formed the sample had the potential to, and perhaps will one day, make strategic acquisition decisions for organizations. They were not very familiar with minibuses, but had seen them, sometimes driven in them, and generally were familiar enough with them to make a weight assessment concerning two attributes that they should be able to comprehend to a certain extent. Based on the interviews afterwards, we have reasons to believe that we succeeded in this respect.

It is by no means certain that the constraints of this study impair the applicability of our findings to real-life decision processes. Having thought about the weights and hence having a frame of reference to work with, actors may adjust their judgements in later stages of the decision process, for example when they actually have to choose between alternatives or when they have to communicate with other actors. This seems especially likely with non-routine decisions, the kind of decisions this study is about. So, the importance assessments made during the preliminary stage of decision processes may well be similar to those observed in our study. Also, in some elicitation methods, actors are asked to give weights at an ordinal level (is attribute A more or less important than attribute B). The weight judgements the subjects made in our study would likely be valid enough to be used in these elicitation methods. And we should remember that, contrary to the solving of logical, well-structured problems, there are no 'right' or 'wrong' weight judgements.

Now that we know in broad terms how people make importance assessments when they are influenced as little as possible by factors like prior experience and the organizational context, we can progressively introduce these factors and observe their influence on the weight assessment process. We conclude this article by stating some expectations concerning the behaviour of experts performing importance assessments and by suggesting some issues for future research.

Behaviour of experts

The results of our study can be interpreted in the light of research on differences on the behaviour of layman and experts. Important characteristics of experts are that they have a large amount of content knowledge and that they are capable of representing a problem in the abstract terms needed for the application of general solution algorithms or heuristics [2]. We therefore expect experts, even in non-routine decision-making:

- to use ready-made or self-developed formal definitions of attributes, aimed at finding characteristics for which a weighting algorithm or heuristic can be applied. This may include an absolute importance scale;
- when decomposing attributes and intending to assign weights to the resulting sub-attributes, to pay attention to the system of decomposition so as to make it complete but not redundant (regardless of how successful the expert is in this respect);
- to integrate sub-attributes to the maximum extent so as to reduce the number of (sub-)attributes to be weighted and to find common denominators for as many (sub-)attributes as possible;
- to devote only a limited amount of effort to evaluation.

The first two skills certainly can be developed. Concepts like mental models or cognitive maps seem suitable for this. So, although these instruments have yet to be adapted for developing importance assessment skills, our research so far shows that the importance assessment process can be analysed, understood and possibly improved (by the use of tools like cognitive mapping).

It is easy to understand that in the case of the minibuses experiments the students could not fall back onto a large arsenal of domain-specific knowledge that would have enabled them to systematically decompose and integrate sub-attributes that are important predictors of successful decision processes. This is why their behaviour is to a large extent hindered compared to experts who do have a huge amount of domain-specific knowledge. The subjects did not establish causal relationships between sub-attributes. It may be likely that they lacked the solution methods (for example, constructing a mental model) that experts are likely to have.

Would experts do better than our subjects in performing importance assessments? This question cannot be answered on the basis of this study. Note that no explicit comparison is made between laymen and experts, as experts formed no part of the research population. Our aim was first and foremost to investigate the behaviour of laymen. In future research, comparisons with experts can be made. Experts would, because of their content knowledge, almost certainly have explicit or implicit definitions of (sub-) attributes available, and they would have a clearer view of the causal relationships between (sub-) attributes. They could, for example, realize that the weight of a minibus determines the braking distance but is in itself not important for safety and hence can be ignored. If the subjects in our study had tried to assess the causal relationships between the numerous sub-attributes they generated, they would probably have been able to eliminate many of them. De Boer [35] gives an overview of methods for establishing causal relationships in an acquisition decision context. It is interesting however, that our subjects, while obviously having less content knowledge than experts, made no attempt to use the content knowledge available to them to assess causal relationships. They simply were not interested in these relationships. So, if there is a difference between laymen and experts in the general way of assessing weights, it may not be the content knowledge as such but *the use* of the content knowledge available.

The question then becomes: do experts pay more attention to, for example, defining attributes instead of just listing sub-attributes, assessing causal relationships and integrating sub-attributes? One would expect so, since experts have more solution methods available and are better in making abstract representations of problems. But on the other hand, it is distinctly possible that experts, even when confronted with non-routine problems, rely on past weight judgements, especially in the early stages of a decision process. So, in our example, the manager of the minibus company may remember why he bought the minibuses he owns at present, and might ask himself what has changed since he bought them. Experts may also, even when they make an individual weight judgment and before communicating it with others, take the organizational context they work in into account. Our subjects certainly wondered what was best for the company they were asked to imagine working for, but they simply lacked information about the organizational content and probably would be hard-pressed to imagine its real-life consequences anyway. Our study certainly provides the basis for formulating hypothesis in future research.

So, whereas our study points at possible differences between laymen and experts, research in real-life decisions is needed to determine which of these differences manifest themselves in practice.

Recommendations for further research

We would like to suggest the following areas of research. Firstly, research on importance assessment by experts. Are they more rational, systematic or consistent than laymen? Do they go through the same phases in much the same way, but only using more content knowledge, or is their approach fundamentally different? For example, do they attempt to find a common denominator for comparing attributes? Another interesting question is whether experts are better in making importance judgements outside their own expertise area. Some authors have suggested that expertise is confined to a certain area; experts are not better than laymen in solving problems outside their own area of expertise. By comparing the way experts perform weight assessments both within and outside their area of expertise we can find out whether this limitation on expertise also holds for a higher-order problem-solving skill like performing weight assessments. This research should concern a range of weighting problems (minibuses, organizational strategies, hiring of personnel), with various numbers of attributes. Using the coding scheme and analytical tools that we developed and which are available to others on demand could do part of this research. It could provide a basis for statistically valid quantitative analysis, which was not possible with the limited number of subjects in our study. It might even be possible to set up a database with think-aloud protocols for this purpose.

Secondly, one could conduct research as a basis for developing and testing instruments aimed at helping actors who have to make weight assessments, for example by stimulating them to decompose attributes systematically or take (causal) relationships between attributes into consideration. An important pitfall here is that the effectiveness of these instruments is difficult to measure. A decision-maker may not know for years whether he made the right decision. For lack of comparison, he may never know. Indicators of the effectiveness of weight assessments may be the perceived quality of the assessment by a panel of experts, or the converging of experts' weight assessments after they have been taught to use the instruments, compared to the situation before the use of the instruments was taught.

Finally, research into other areas of management should be conducted. Management science is a multidisciplinary field. Heuristics have been developed for integrating information from several disciplines into a decision on, for example, the strategy of a company (see for an overview [35]). But what goes on in the head of managers? Do they really follow these normative heuristics? Insofar as integrating information from various disciplines means assessing the weight of factors to be taken into consideration, our model may be used for analysing the reasoning of managers in different fields.

References and notes

1. Van der Heijden, B.I.J.M. (1998). *The measurement and development of professional expertise throughout the career. A retrospective study among higher level Dutch professionals*. Ph.D.-thesis. University of Twente, Enschede, the Netherlands. Enschede: PrintPartners Ipskamp.
2. Chi, M.T.H., Glaser, R., & Farr, M.J. (Eds.) (1988). *The nature of expertise*. Hillsdale, NJ: Lawrence Erlbaum Associates
3. Simon, H.A (1979). Information Processing Theory of Human Problem Solving. In: Estes, W (ed.) *Handbook of Learning and Cognitive Processes* (Vol. 5). Hillsdale: Lawrence Erlbaum Associates Inc.
4. Keeney, R.L., & Raiffa, H. (1976). *Decisions with multiple objectives, preferences and value tradeoffs*. New York [etc.]: John Wiley & Sons.
5. Bereiter, C., & Scardamalia, M. (1993). *Surpassing ourselves. An inquiry into the nature and implications of expertise*. Chicago: Open Court.
6. Ericsson, K.A. (1996). *The road to excellence. The acquisition of expert performance in the arts and sciences, sports, and games*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.
7. Ericsson, K.A., & Smith, J. (1991). *Toward a general theory of expertise. Prospects and limits*. Cambridge: Cambridge University Press.
8. Johnson, E.J. (1988). Expertise in admissions judgment. Unpublished doctoral dissertation. Carnegie-Mellon University.
9. Camerer, C.F., & Johnson, E.J. (1991). The process-performance paradox in expert judgment. How can experts know so much and predict so badly? In: Ericsson, K.A., & J. Smith (Eds.), *Toward a general theory of expertise. Prospects and limits*. Cambridge: Cambridge University Press.
10. Patel, V.L., & Groen, G.J. (1991). The general and specific nature of medical expertise: a critical look. In: Ericsson, K.A., & J. Smith (Eds.), *Toward a general theory of expertise. Prospects and limits*. Cambridge: Cambridge University Press.
11. Dam, K. van (1993). Beoordelingsprocessen: structuur- versus procesgericht onderzoek. In: Koele, P. & van der Pligt, J. (Ed) *Beslisten en beoordelen*. Meppel: Boom
12. Van der Heijden, B.I.J.M. (2000). The development and psychometric evaluation of a multi-dimensional measurement instrument of professional expertise. *High Ability Studies. The Journal of the European Council for High Ability*, 11(1), 9-39.
13. Sternberg, R.J. (1995) Expertise in complex problem solving. A comparison of alternative conceptions. In: P.A. Frensch & J. Funke (ed.) *Complex problem solving. The European perspective* (pp. 295-321). Hillsdale NJ: Lawrence Erlbaum Associates, Publishers
14. Glaser, R. (1987). Thoughts on expertise. In: C. Schooler & W. Schaie (Eds.), *Cognitive functioning and social structure over the life course* (pp. 81-94). Norwood, NJ: Ablex.
15. Curtis, B. (1986). By the way, did anyone study any real programmers? In: E. Soloway & S. Iyengar (Eds.), *Empirical studies of programmers*. Norwood: New Jersey: Ablex.
16. Logan, G.D. (1985). Skill, and automaticity: relations, implications and future directions. *Canadian Journal of Psychology*, 39, 367-386.
17. McLagan, P.A. (1997). Competencies: the next generation. *Training & Development*, 51(5), 40-47.
18. Ericsson, K.A., & Lehmann, A.C. (1996). Expert and exceptional performance: Evidence on maximal adaptations on task constraints. *Annual Review of Psychology*, 47, 273-305.
19. Trost, G. (1993). Prediction of excellence in school, university and work. In: K. Heller, F.J. Mönks & A.H. Passow (Eds.), *International Handbook of Research and Development of Giftedness and Talent* (pp. 325-336). Oxford: Pergamon Press.
20. Alexander, P.A., Schallert, D.L., & Hare, V.C. (1991). Coming to terms: How researchers in learning and literacy talk about knowledge. *Review of Educational Research*, 61(3), 315-344.
21. Van der Heijden, B.I.J.M. (1996). *Life-long expertise development: goal of the nineties*. Book of Proceedings of the Fifth Conference on International Human Resource Management. Hyatt Islandia San Diego, California, USA, 24-28 June 1996.

22. Tikkanen, T. (1998). *Learning and education of older workers: Lifelong learning at the margin*. Jyväskylä Studies in Education, Psychology and Social Research 137. University of Jyväskylä.
23. Evans, J.R. (1991). *Creative thinking in the decision and management sciences*. Cincinnati: South-Western Publishing Co.
24. Newell, A., & Simon, H.A. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall
25. Ericsson, K.A., & Simon, H.A. (1993). *Protocol analysis: verbal reports as data* (2nd ed.). Cambridge, MA [etc.]: Bradford Books/MIT Press.
26. Simon, H.A. (1976). *Administrative behaviour*. New York: The Free Press, London: Collier McMillan Publishing.
27. Kotovsky, K. & Fallside, D. (1989). Representation and transfer in problem solving. In: Klad, D. & Kotovsky, K. (ed). *Complex information processing; the impact of Herbert A. Simon*. Hillsdale, NJ, 69-108
28. Strauss, A., & Corbin, J. (1998). *Basics of qualitative research; Grounded Theory, procedures and techniques*: Thousand Oaks CA [etc.]: Sage Publications.
29. Baarda, D.B. & de Goede, M.P.M. (2001). *Basisboek methoden en technieken*. Groningen: Stenfert Kroese.
30. Hicks, M.J. (1995). *Problem solving in business and management; hard, soft and creative approaches*. London [etc.]: International Thomson Business Press.
31. Akin, Ö. (1994). Creativity in design. *Performance Improvement Quarterly*, volume 7, 9-23.
32. Timmermans, D. (1993) The impact of task complexity on information use in multi-attribute decision-making, *Journal of behavioural decision making*, 6, 95-111.
33. Kahneman, D, Knetch, J.L. & Thaler, R.H. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of political economy*, volume 96, 1325-1348.
34. Kahneman, D. (1994) New challenges to rationality assumption, *Journal of institutional and theoretical economics*, volume 150, 18-36
35. Boer, L. de (1998). *Operations research in support of purchasing; design of a toolbox for supplier selection*. Enschede: Universiteitsdrukkerij.
36. Johnson, G, & Scholes, K. (1999). *Exploring corporate strategy, fifth edition*. London [etc]: Prentice Hall.

DE VOLGENDE REFERENTIES ZATEN IN DE OORSPRONKELIJKE VERSIE MAAR ZIJN ER LATER UITGEVALLEN

37. Harte, J.M., & Koele, P. (1995). A comparison of different methods for the elicitation of attribute weights: structural modelling, process tracing and self-reports. *Organizational Behaviour and Human Decision Processes*, 64, 49-64.
38. Aronstein, D.C, Hirschberg, M.J., & Piccirillo, A.C. (1998). *Advanced Tactical Fighter to F-22 Raptor*. Reston: American Institute of Aeronautics and Astronautics Inc.
39. Enserink, B. (1993). *Influencing military technological innovation*. Enschede: PhD thesis.
40. Borcherdig, K, Schmeer, S., & Weber, M. (1995). Biases in multi-attribute weight elicitation. *Contributions to decision making-I*, Caverni, J.P, Bar-Hillel, M. & Jungemann, H. (ed). Amsterdam [etc.]: Elsevier Science BV.
41. Keeney, R.L (1992). *Value-focused thinking*. Cambridge MA [etc.]: Harvard University Press.
42. Anderson, J.C, Thomson, J.B.L., & Wynstra, F. (2000). Combining value and price to make purchase decisions in business markets. *International Journal of Research in Marketing*, 17, 307-329.
43. Meehl, P.E. (1954). *Clinical versus statistical prediction*. Minneapolis: University of Minnesota Press.
44. Hunt, E. (1994). Problem solving. In: Sternberg, R.J. (Ed) *Thinking and problem solving*. San Diego [etc.]: Academic Press.
45. Aumann, R. (1997). Rationality and bounded rationality. *Games and Economic Behaviour*, volume 21, 2-14.
46. Chapman, G.B., & Johnson, Eric. J. (1999). Anchoring, activation and the construction of values. *Organizational Behaviour and Human Decision Making*, 79, 115-153.

47. Wong, K.F.E. & Kwong, J.Y.Y. (2000). Is 7300m equal to 7.3 km? Same semantics but different anchoring effects. *Organizational Behaviour and Human Decision Processes*, volume 82, 314-333.
48. Cascio, W.F. (1991). *Applied psychology in personnel management*. London: Prentice-Hall International.

Figures

Table 1. The phases of the WAM

Phase number	Phase name	Percentage of segments devoted to the phase	Percentage of subjects with which elements of phase were observed
1	Problem identification	6.74 %	100 %
2	(Sub-) attribute processing	30.33 %	100 %
3	Absolute sub-attribute weighting	27.22 %	100 %
4	Homogeneous sub-attribute weighting	4.53 %	66.7 %
5	Heterogeneous sub-attribute weighting	1.50 %	55.6 %
6	Attribute weighting	12.54 %	100%
7	Evaluation	17.14 %	100%

Table 2. The decomposition of ‘safety’

Number of attributes as a result of decomposition	Number (%) of subjects, first level	Number (%) of subjects, second level	Number (%) of subjects, third level
0		4 (22)	15 (83)
1-5	2 (11)	7 (39)	1 (6)
6-10	5 (28)	4 (22)	1 (6)
11-15	5 (28)	2 (11)	1 (6)
16-20	5 (28)	1 (6)	
21-25			
26-30			
30-35	1 (6)		

Table 3. Integration of safety and comfort

Number of sub-attributes being integrated	Number (%) of subjects integrating sub-attributes of safety	Number (%) of subjects integrating sub-attributes of comfort
0	9 (50)	13 (72)
1-5	6 (33)	5 (28)
6-10	3 (17)	