Thesis number: INF/SCR-09-39

Supporting computer-aided decision making by improving alignment between human experts and probability elicitation instruments

PERSONAL AND SUPERVISOR'S INFORMATION

Personal
Name: Gwyneth Ouwehand
St.nr: 0371742
Master: Content and Knowledge Engineering
E-mail: g.ouwehand@gmail.com

1st Supervisor
Name: Dr. Ir. R.J. Beun
Group: Cognition and Communication
Section: Game, Media and Agent Technology
E-mail: rj@cs.uu.nl

2nd Supervisor
Name: Dr. S. Renooij
Group: Decision Support Systems
Section: Information and Software Systems
E-mail: silja@cs.uu.nl

Universiteit Utrecht

Abstract

The purpose of this thesis was to provide recommendations for better alignment between human experts and probability elicitation methods. In order to contribute to this goal one probability elicitation method in particular was investigated: the verbal-numerical scale developed by Renooij and Witteman (1999). Two experiments were conducted. Experiment 1 was a qualitative think aloud experiment in which experienced veterinarians answered probability questions while thinking aloud. The goal of this experiment was to gain insight in the cognitive processes of answering probability questions and the use of the verbal-numerical scale. The subjects indicated their answers to the probability questions on the original verbal-numerical scale or on an alternative verbal-numerical scale that was designed for this experiment. It was observed that both scales (original and alternative) attracted different use of it. To find the cause of this difference a second (pilot) experiment was set up that investigated the basics of how both scales are perceived and interpreted using the theory of gestalt principles. This pilot experiment presented subjects with several images of both the original and the alternative representation of the verbal-numerical scale. Subjects were asked to cluster parts of the scale which they thought belong together by encircling them. The cause of the different behavior on the two scales was not found. It was however shown that semantics are the main reason for clustering objects and not the form and position of objects. The principal conclusion was that the current representation of the verbal-numerical scale is causing a problem with the reliability of the answers indicated on the scale. Several recommendations are given about feedback in probability elicitation methods, adapting to the information need of the experts, the use of probability words and probability numbers in the verbal-numerical scale.

Contents

Chapter 1: Introduction

Every day we make hundreds of decisions. What do we eat for breakfast? Should we take an umbrella if we go outside? What t-shirt shall I wear today? These decisions are relatively easy. If the wrong decision is made, the consequences are not really serious, and the amount of information we have to take into consideration can be overseen. Some decisions are more difficult, like a doctor who has to decide whether to operate on a patient or not or a veterinarian that has to make a diagnosis on a sick pig. A decision can be difficult when a lot of information needs to be considered, when the consequences of a decision are serious, or when the decision maker does not have enough knowledge or experience to make the decision [1]. Nowadays difficult decisions are often supported by so called decision support systems (DSS).

1.1. Decision Support Systems and Bayesian Networks

A DSS can provide a decision maker, for instance a doctor, an overview with the possible diagnoses and treatments for a certain patient and indicate what the consequences are of choosing a specific treatment. In a DSS several components can be distinguished: an user interface, a database management system (DBMS) and a model–based management system (MBMS) [2]. Through the user interface the user can access the data and models that are contained in the DBMS and MBMS. The DBMS contains data needed for the decisions under consideration, like a list of diagnoses, medications, diseases and treatments. The MBMS contains models in which is described how these data are related. Several kinds of models exist and they describe a certain (problem) domain [2].

In this thesis we will focus on a method for the collection of data for a specific kind of model in a MBMS: a Bayesian Network. In short, a Bayesian network is a mathematical model that enables reasoning with uncertainty to support decision making, but cannot make decisions of its own. We focus on a Bayesian Network that models the domain of Classical Swine Fever (CSF). Bayesian Networks represent a joint probability distribution and require numerical probabilities for their specification. Since there are no sufficient literature and data sets available about CSF, these numerical probabilities need to be elicited from human experts, in this case experienced veterinarians. To elicit numerical probabilities from experts, several probability elicitation methods have been developed, like numerical scales and probability wheel methods (for an overview see [3]).

1.2. Verbal-numerical scale

For a typical Bayesian Network thousands of probabilities need to be collected. Therefore a probability elicitation method is needed that collects probabilities from experts at a high speed. If the model is completely specified, sensitivity analysis can point out what probabilities are important. These probabilities can be collected more accurately with a more time consuming elicitation method. The existing probability elicitation methods are time consuming or do not provide adequate support for the expert. In [4], a verbal-numerical scale was developed in order to speed up the elicitation process and meanwhile support the expert in providing a numerical probability. This scale was subsequently put to use by Van der Gaag et al [5] as part of a new probability elicitation method. This method consisted of a verbal description of the required probability along with the scale depicted in Figure 2.3. This verbal-numerical scale has already been used extensively, both in The Netherlands and in other countries. For instance, Van der Gaag et al [5] used it to elicit probabilities for a system that was developed for patient-specific therapy selection for esophageal carcinoma, Geenen et al [6] for a system that detects classical swine fever, and Charitos et al [7] for diagnosis and treatment of ventilator-associated pneumonia. The scale is available in both Dutch and English.

1.2.1. Problems

An important caveat here is that, at least locally, the method has always been used in guided one-on-one elicitation sessions of the domain expert with an elicitor familiar with the method, in a native-Dutch setting. There were some problems identified in these sessions, like experts who indicated their answer on the scale in a way that was not intended. These problems during the sessions never lead to problems in the results, because the elicitor was able to correct the expert during the interactive one-on-one session. Recently the method was used in an unguided group setting instead. The scale was used in sessions with small groups of (five to seven) veterinary experts in the six partner countries of the so called EPIZONE project. Most of the domain experts were non-native English or Dutch speakers. The experts reportedly did speak the English or Dutch language[8]. They were informed about the intended use of the verbal-numerical scale and instructed in a plenary introduction to indicate their answer as a horizontal dash or a cross on the vertical line of the scale [9]. In this unguided group setting the individual effort of the experts could not be traced and the answers were analyzed after the session without the presence of the expert, instead of during the session in conjunction with the expert. The analysis of the answers of the experts after the sessions and some observations made during the sessions, lead us to believe that there are some problems with the use of the verbal-numerical scale. The observations and problems are extensively described in chapter 3. Here, the observations are shortly listed:

- Experts got frustrated and walked away without finishing the task
- Experts indicated on the scale how sure they were of the answer instead of the answer to the question
- Experts indicated their answer on the scale in different manners than instructed
- Some probability answers of the experts did not add up to one as they should have
- More than half of all answers were indicated on anchors on the scale, being a verbal label, a numerical label or a horizontal dash.

These observations lead to problems with the interpretation of the answers and the reliability of the answers. To conclude, despite of successful use of the verbal-numerical elicitation method in guided one-on-one sessions, the method seems to have some problems when used in an unguided group setting.

1.3. Goal

The goal of this thesis is to provide recommendations for better alignment between human experts and probability elicitation methods like the verbal-numerical scale. We aim to contribute to this goal by investigating one probability elicitation method in particular: the verbal-numerical scale. We conducted two experiments for this thesis. In Experiment 1 we studied experienced veterinarians as our subjects while they were answering probability questions while thinking aloud. The subjects indicated their answers on the original verbal-numerical scale, as described above, and an alternative verbal-numerical scale we designed for this experiment. It was observed that both scales (original and alternative) attracted different use of it. To investigate this difference we set up a second experiment that investigated the basics of how both scales are perceived and interpreted.

1.4. Thesis overview

This thesis is organized as follows. In Chapter 2 the main concepts of this thesis are introduced. The use and construction of Bayesian models is explained and the available probability elicitation methods are described. Furthermore the cognitive processes of answering a question are outlined. In Chapter 3 we discuss the development of the verbal-numerical scale in more detail and show the specific problems that occurred with the use of the scale, resulting in our hypotheses about the representation of the scale and the method of the scale. Chapter 4 contains the experimental design of the first experiment; a qualitative think aloud experiment to gain information about the thoughts and motivation of subjects while answering probability questions using two representations of the verbal-numerical scale. The results of this experiment are presented in Chapter 5. Chapter 6 evaluates the hypotheses with the results from Experiment 1. Some questions remain unanswered, which leads to Experiment 2. The design of this pilot experiment is described in Chapter 7. The results of this experiment are presented in chapter 8. The last chapter contains the conclusion, some recommendations and ideas for further research.

Chapter 2: Bayesian networks, probability elicitation methods and cognitive processes

This chapter describes the use and construction of Bayesian networks, provides a short overview of available probability elicitation methods and outlines the cognitive processes of question answering. In the first part it is outlined what kind of data is required for the specification of a Bayesian network, namely a graph that captures variables and their inter dependencies, and probabilities in a numerical format. The second part of this chapter describes the different probability elicitation methods designed to elicit numerical probabilities from experts, including the verbal-numerical scale. We briefly list the problems that can occur when people estimate probabilities. The third and last part discusses the steps that occur while people are answering questions.

2.1. Bayesian networks

This section starts with an explanation of Bayesian networks. Section 2.1.2 describes the Bayesian network that has been developed to model the domain of Classical Swine Fever (CSF). The last section outlines the data that is needed for the construction of a Bayesian network.

2.1.1. Definition

A Bayesian network is a mathematical model that provides for reasoning under uncertainty. More specifically, a Bayesian network is a concise representation of a joint probability distribution over a set of variables. It combines a graphical representation that captures the inter dependencies among the variables, with conditional probability mass functions.

Bayesian networks were first introduced by Pearl in the 1980s. The networks use efficient algorithms to enable computing any probability distribution of interest for one or more of its variables. To this end, the algorithms basically apply several rules from probability theory, among what is called Bayes' rule, named after the British mathematician Thomas Bayes who lived in the 18th century. He introduced a mathematical rule that relates conditional and marginal probabilities of two events, which is often used to compute posterior probabilities given observations [10]. The conditional probability of A being true given that B is true is mathematically notated as P(A|B) and can be

established from the conditional probability P(B|A), and the marginal probabilities P(A) and P(B) using Bayes' rule:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

2.1.2. Bayesian network for CSF

In Figure 2.1. the Bayesian network for the domain of early detection of Classical Swine Fever (CSF) is depicted. It represents the probabilistic relationships between various symptoms and between those symptoms and the disease CSF. Given some symptoms the network can calculate the probability that CSF is present. In order to reason with the Bayesian network it needs to be completely specified, that is, all relevant variables and their inter dependencies are modeled in an acyclic directed graph, where the nodes (circles) are the variables and the edges (arrows) are possible dependencies. In addition, all required conditional probabilities, as is described below, should be specified in their respective tables.

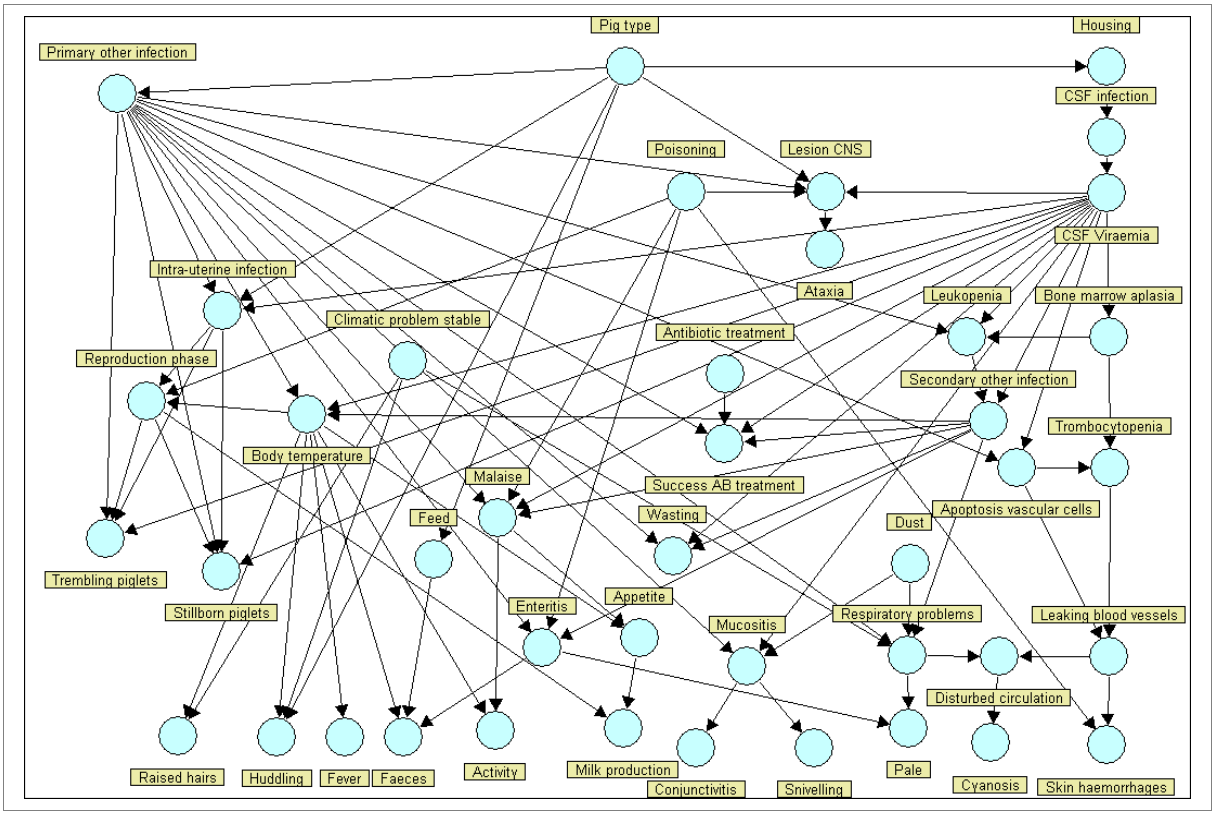


Figure 2.1. Graphical representation of the CSF network

Examples of variables in the CSF-network are for instance 'Body temperature', 'Malaise' and 'Appetite'. The directed relations indicate that the variable 'Appetite' is dependent of both the variables 'Body temperature' and 'Malaise'. All variables have a finite number of mutually exclusive and collectively exhaustive values; the appetite' can be normal or reduced, the body temperature can be normal or elevated and the malaise can be present (yes) or not (no) (See Figure 2.2).

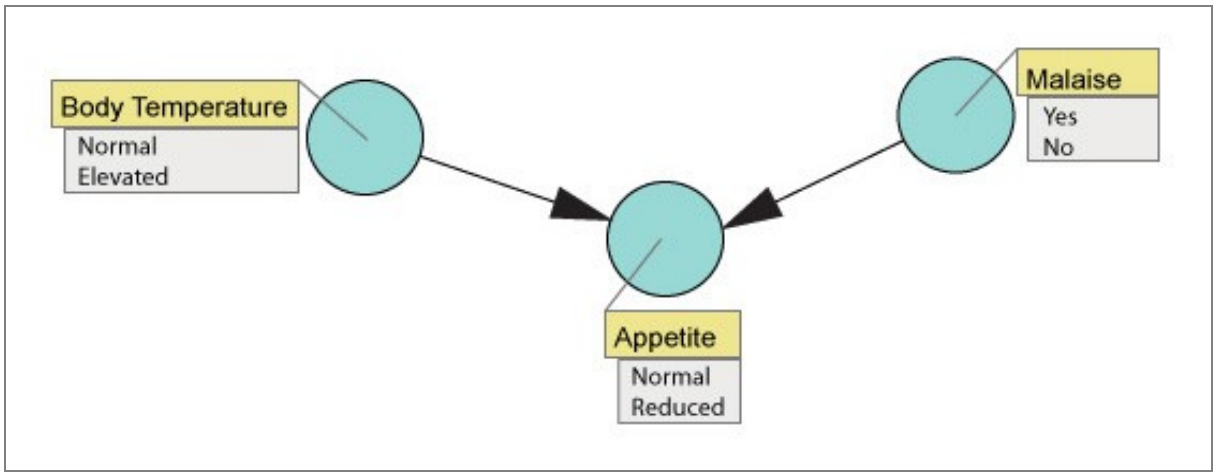


Figure 2.2. Detail of the Bayesian Network for CSF

With each variable is associated a probability table describing the strength of the effect, in terms of a probability of each combination of values for the parents of the variable on each value of the variable itself. For example the probability of a reduced appetite is influenced by the values of the variables 'Body temperature' and 'Malaise'. The probability table of the variable 'Appetite' provides a probability for each of its values, given every combination of values from the variables of its parents 'Body temperature' and 'Malaise' in the graph (see Table 2.1.).

Malaise	*No*		*Yes*	
Body temperature	*Normal*	*Elevated*	*Normal*	*Elevated*
Appetite *Normal*	0.995	0.75	0.15	0.1
Appetite *Reduced*	0.005	0.25	0.85	0.9

Table 2.1. Probability table of the variable Appetite

As can be seen in Table 2.1., the probability that a reduced appetite is caused by a normal body temperature and absence of malaise is 0.005. An elevated temperature in combination with the presence of malaise is more likely to cause a reduced appetite with a probability of 0.9.

When the network is completely specified, then any (conditional) probability of interest can be calculated. We can for instance enter a reduced appetite as observation, upon which the reasoning algorithms can calculate the probability that the body temperature is elevated. More importantly, the probability can be calculated that the observed symptoms are caused by CSF.

2.1.3. Data for constructing Bayesian Networks

To construct a Bayesian network two types of data are required:

1. Variables, dependencies between those variables and values of the variables.
2. Probabilities for every combination of the values of the variable and of all the variables it directly depends on.

As mentioned before, this data is usually collected using literature, databases with cases (data sets), and/or human experts. In the domain of CSF, however, the literature cannot provide the probabilities that are necessary and databases with cases were not available to this end. The collection of

probabilities for this network totally depended on the judgment of human experts. The data about the variables, values and dependencies for the construction of the CSF network was collected by interviewing experts in the field. The Bayesian network for CSF includes 42 variables and requires 2400 probabilities. To gather such big amounts of probabilities takes a lot of time, while experts usually do not have much time. The probabilities need to be entered in the network's specification as numerical values. Experts usually have minimal mathematical knowledge. Therefore, experts need to be supported in answering probability questions in a numerical probabilistic format. For this purpose several elicitation methods have been developed.

2.2. Probability elicitation methods

Since humans are not good in estimating probabilities [11], several probability elicitation methods have been developed. Examples are probability-wheel and gamble-like methods and probability scales (for an overview see [3]).

The probability-wheel and gamble-like methods are time consuming. Furthermore they are difficult to learn because the experts need to understand the underlying mathematical concepts. For instance, before an expert can use a gamble-like method he has to understand the concept of decision trees.

As a direct method to elicit probabilities, a probability scale is often used. The underlying idea is to support experts in expressing their estimation by thinking of visual proportions instead of a precise number. Various probability scales exist, varying in the amount of numerical labels and their positions. Probability scales are considered easy to understand and easy to use. The sessions with the experts are not time consuming and therefore suitable for the elicitation of probabilities from experts.

In a case study of Van der Gaag et al [5] it was found that expert oncologists felt uncomfortable working with a probability scale with only numerical anchors; it gave them 'very little to go by'. Therefore Renooij and Witteman [4] developed a scale with both verbal labels and numerical labels; the verbal-numerical probability scale (Figure 2.3.) on which this thesis focuses. They chose to depict verbal labels in addition to only numerical labels, based on the theory that people prefer to express and process probabilities in a verbal rather than a numerical form [12]. Therefore they thought the verbal labels would help the experts express their estimation.

2.2.1. Design and use of the verbal-numerical scale

Renooij and Witteman figured that a double scale, with both numerical and verbal anchors, might be helpful since experts prefer to communicate their probability assessment in a verbal format. A series of four studies resulted in the current design of the verbal-numerical scale. For more details on these studies, see Renooij and Witteman [4]. The studies resulted in a list of seven commonly used verbal expressions, their rank order and the numerical probability they could be projected on: certain 100%, probable 85%, expected 75%, fifty-fifty 50%, uncertain 25% and impossible 0%.

The verbal-numerical scale was designed to contain these seven verbal labels and seven numerical labels (Figure 2.3.) The verbal labels were placed on the left side of a vertical line with short horizontal dashes on it, and the numerical labels were placed on the right side of the vertical line. The dashes on the vertical line are positioned at the same height as the numerical labels, to express that they belong together. The verbal expressions are not directly positioned next to the numerical labels; they are distributed more evenly over the scale, closer to the center, to prevent the verbal labels incorrectly to be taken as exact translations of precise numbers. The scale is meant to be perceived as a set of labels with a stable rank ordering covering the whole probability continuum. The scale is intended to be continuous and to allow subjects to indicate any degree of probability. The word 'almost' is added to the first and last verbal label (certain and impossible) to indicate the positions of very small and very large probabilities.

Subjects are supposed to mark their probability estimate as a cross or line on the vertical line of the verbal-numerical scale, providing the elicitor a point estimate. The answer is interpreted by measuring the distance between the 0 and the mark on the vertical line.

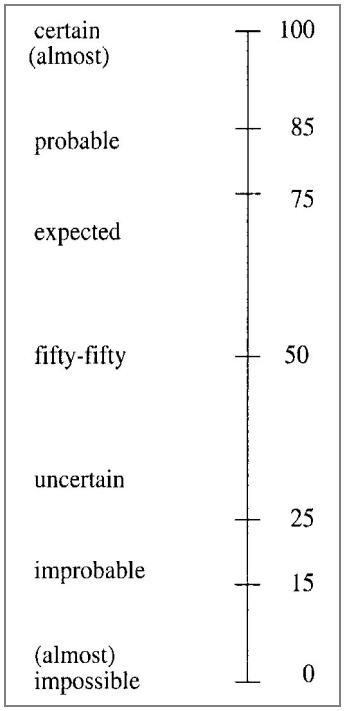


Figure 2.3. Verbal-numerical probability scale by Van der Gaag et al[1]

This section described the design of the verbal-numerical scale, but the associated elicitation method is more than the scale itself. The context in which the scale is used is also important. As a probability question is posed to the expert, the expert has to find an answer to the question and indicate it on the verbal-numerical scale. This process of answering questions is outlined in more detail in the next section.

[1] The Dutch translation of the verbal labels are "zeker (bijna)", "waarschijnlijk", "te verwachten", "fifty-fifty", "onzeker", "onwaarschijnlijk" and "(bijna) onmogelijk".

2.3. Cognitive processes of answering questions

Sudman et al. [13] describe four steps in the cognitive processes of answering a question.

1. Interpreting the question
2. Retrieving information
3. Generating a response
4. Formatting a response

As a start, the respondent needs to interpret the question. This includes processing the words to understand what kind of information is asked for. Next, the respondent will try to remember relevant instances to the question and assess whether each instance is useful for the answer. From the useful instances, an answer is generated. When the respondent has formulated an answer for himself, this answer needs to be communicated. Therefore the answer should be formatted as is asked in the question, for instance fit the own answer to one of the presented multiple choice answers or map the own answer to a response scale.

Sudman et al. end the steps with the formatting of the response. When using the scale, however, putting down the answer on the verbal-numerical scale is also an important step. We will add this step to Sudman's steps.

5. Indicate the response on the verbal-numerical scale

In order to succeed in getting the answer to the question that was posed all steps should go smoothly.

The next chapter analyzes the problem with the verbal-numerical scale by describing the observations that were made during and after the use of the verbal-numerical scale in the unguided group sessions of the EPIZONE project.

Chapter 3: Problem analysis and hypotheses

In this chapter the problems with the scale during the EPIZONE project are analyzed in detail. In section 3.1. the setting of the unguided group sessions of the EPIZONE project is described. The following section provides an overview of some observations during and after these sessions. In section 3.3 it was assessed to what problems these observations lead for the use of the scale. In section 3.4. the causes of the problems are hypothesized.

3.1. Sessions in group setting

Between December 2006 and May 2007 the verbal-numerical scale was used in sessions with small groups of (five to seven) veterinary experts in the six partner countries of the EPIZONE project: Belgium, Denmark, Germany, Great Britain, Italy and Poland. In the group setting an elicitor (Van der Gaag) lead a group of domain experts and started with a plenary introduction slide show. The meaning and interpretation of the verbal-numerical scale was explained and it was described how to translate the numbers or percentages to frequencies. The presentation contained an example of how to indicate an answer on the scale. This image showed a red dash on the vertical line of the scale. It was mentioned that one also could write a number down with this dash. This was also shown in an image.

The attending experts were each presented with assessments for a number of probabilities in the domain of CSF. The paper answer forms presented the assessment as a text fragment containing a requested probability. Along with each text fragment the verbal-numerical scale was presented. On every sheet they were presented with two probability questions from the same probability distribution; the probability some symptom is observed and the probability that the same symptom will not be observed under the same circumstances. The experts were instructed to answer the question they felt most comfortable answering. It was not necessary for the experts to answer both questions. The experts were asked to carefully consider the fragment of text and to indicate his or her assessment for the requested probability by marking the scale with a dash or a cross on the vertical line. Figure 3.2. shows an example of an assessment from one of the session described above.

Unfortunately, when analyzing the answers of the experts, it appeared that some experts answered in a different way than was instructed. This causes problems for the interpretation of the answers. Some other observations were made during the sessions and afterwords. We will now describe these observations.

3.2. Observations

Two kinds of observations were made. The first category of observations were made by the elicitor during the actual elicitation sessions. The second category contains observations made by analyzing the answers of the experts afterwords. The observations in the first category were reported by the attending elicitor and cannot be derived from the filled scales. The other observations are the result of our analysis of the answers on the filled scales.

3.2.1. Reported observations during the sessions

During the elicitation sessions the attending elicitor did some remarkable observations [9]. These observations are outlined below.

Frustration

Some experts said they could not answer the questions posed to them, even with help from the elicitor they could not do it. At least one expert walked away from the session without finishing it.

Certainty instead of answer

Another observation was that some experts indicated on the scale how certain they were about their answer instead of indicating the probability itself. Certainty is then used in a different meaning. In the task, the certainty of the presented relation of variables is asked. Some experts apparently interpreted certainty as certainty in relation with the answer, the expert's certainty about the answer. One expert for instance circled the verbal anchor 'certain', indicating he was 25% certain about his answer.

3.2.2. Observations from analyzing answers

For this thesis, we analyzed the answers from the experts on the answer sheets. Three interesting observations were noticed. First it was found that many experts did not follow the instructions of how to use the verbal-numerical scale. Instead of indicating their answer with a cross or a dash on the vertical line of the scale, they indicated their answer in all kinds of other manners, for instance by drawing circles around verbal labels. Secondly, we noticed that most of the answers were indicated on an anchor. Third, we found that some experts did fill out both questions that were presented on the same page. Since these two questions represent two complementary probabilities from the same probability distribution the answers should add up to a probability of one. In most cases they did not. All three of these observations are described in more detail below.

Wrongly indicated answers

If an answer is indicated in another way than was instructed, we call it wrongly indicated. This does not necessarily mean that the answer is not useful or incorrect, it is just not indicated in the intended way. Many experts did not use the scale as was presented to them in the instruction, see for examples Figures 3.1.a – 3.1.o below. These figures represent all kinds of wrongly indicated answers. Instead of indicating their estimate with a dash or a cross on the vertical line of the scale, experts drew circles around verbal labels (Figure 3.1.a), around numerical labels (Figure 3.1.b), around a dash on the vertical line (Figure 3.1.c), on a part of the vertical line with no existing dash (Figure 3.1.d) or around combinations of labels, dashes and the vertical line (Figure 3.1.e, 3.1.f, 3.1.g, 3.1.h). Some experts did indicate their answer with a cross or dash but they put it next to a numerical label (Figure 3.1.i) or they put a dash on the vertical line with a circle around it (Figure 3.1.j). Although the instruction basically asked for a point estimate, some experts indicated a probability range (Figure 3.1.k and 3.1.l). There were also scales found on which experts encircled a verbal label and a line connecting it to the vertical line (Figure 3.1.m). One expert even drew a circle between two verbal labels and connected that to the vertical line (Figure 3.1.n), and another expert circled a verbal label and also put a dash on the vertical line (Figure 3.1.o).

Answers indicated on anchors

Besides the many wrongly indicated answers, we noticed that more than half of all answers was indicated on an anchor on the scale, that is a verbal label, a numerical label or a dash. We counted all answers that were given on a anchor. The example Figures 3.1.a, b, c, e, and i are for instance all judged to be answers on an anchor. A total of 412 questions were answered in the group sessions. A total of 39 experts participated in the sessions. Around 40% of all questions were answered using the flexibility provided by the continuous scale, that is indicating a point probability that is not on an anchor (see Table 3.1.).

Anchor	*232*	56.3%
Point	*166*	40.3%
Else	*14*	3.4%
Total	***412***	***100%***

Table 3.1. Overview of the amount of answers indicated as an anchor, a point or in another way (like in Figure 3.1.f).

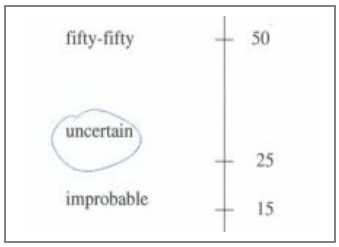


Figure 3.1.a

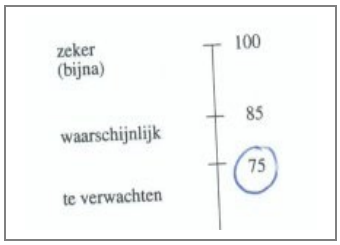


Figure 3.1.b

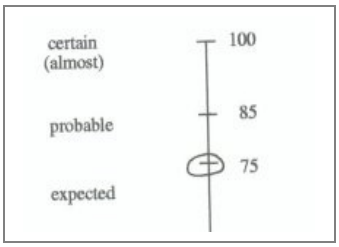


Figure 3.1.c

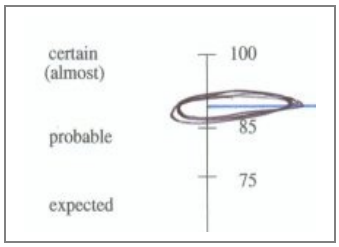


Figure 3.1.d[2]

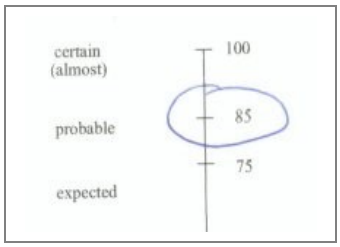


Figure 3.1.e

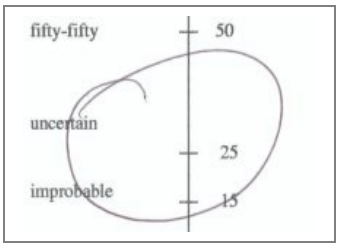


Figure 3.1.f

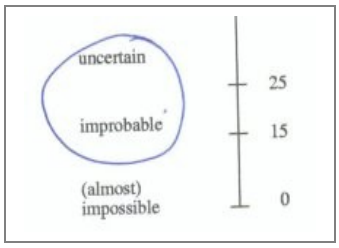


Figure 3.1.g

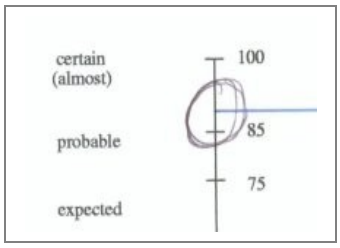


Figure 3.1.h[1]

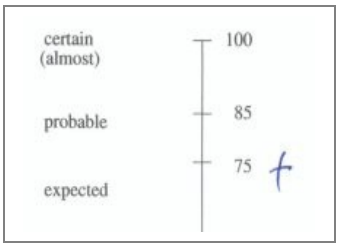


Figure 3.1.i

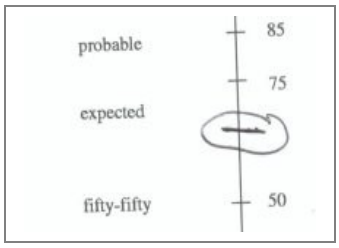


Figure 3.1.j

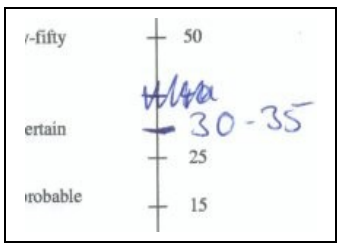


Figure 3.1.k

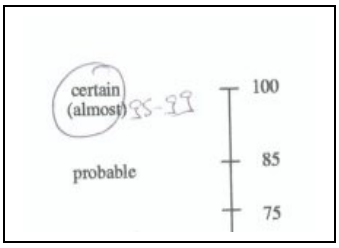


Figure 3.1.l

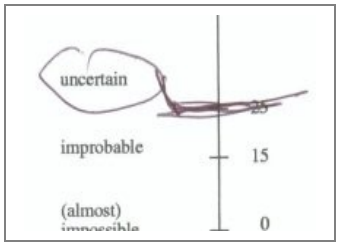


Figure 3.1.m

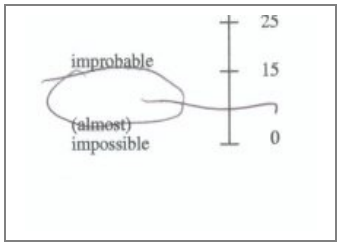


Figure 3.1.n

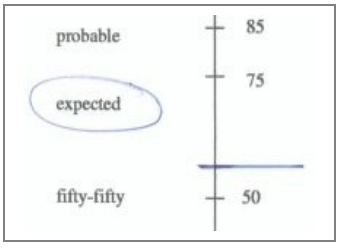


Figure 3.1.o[1]

Figure 3.1. Details from answers on the verbal-numerical scale in the EPIZONE project.

[2] The blue horizontal line should be ignored . This line was drawn by the elicitor in the evaluation, not by the subject during the experiment.

Chances not adding up to one

The probabilities to be estimated were always presented as positive (the chance that something occurs) and negative (the chance that something does not occur) on one page. Only one of which needed to be filled out, as indicated in the introduction. Some experts filled out both parts, but the answers indicated in both did not add up to 1, as it should (see Figure 3.2.).

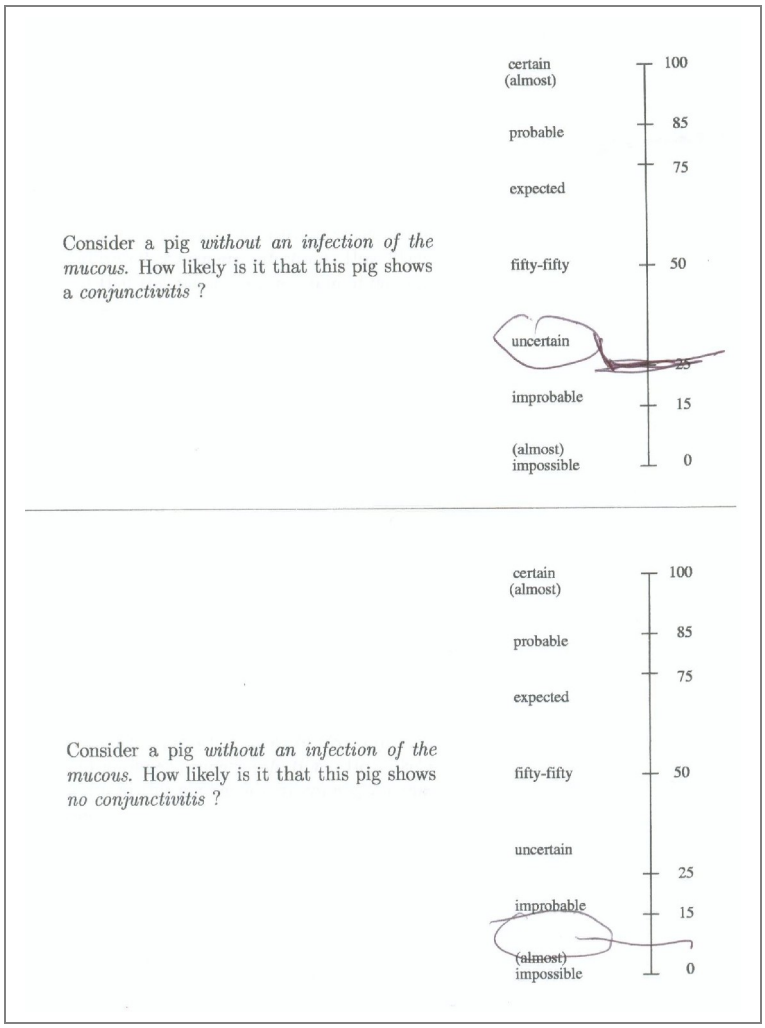


Figure 3.2. Answer sheet from an expert from the EPIZONE project.

3.3. Problems

The observations lead to problems with the use of the verbal-numerical scale. It is clear that experts who got frustrated and walk away cause a problem, because their expertise cannot be captured. Besides that, two main problems are detected. The first problem concerns the interpretation of the wrongly indicated answers. Some answers cannot be interpreted by the elicitor and cannot be used in the Bayesian Network. The second problem is the reliability of the answers. With reliability we mean to which extent the elicitor can rely on his interpreted answers to be the answers to the questions he posed. During the group sessions it was observed that at least one expert indicated how

certain he was of his answer instead of indicating the answer itself. The two problems of interpretation and reliability are described in more detail below.

3.3.1. Interpretation

If answers are indicated in another way than was instructed, it is in some cases hard or even impossible to associate a number with the answers that were indicated. These answer can therefore not be used in the Bayesian Network.

The answers indicated as an encircled verbal expression (like Figure 3.1.a) could not be used for numerical processing since these labels indicate a more or less fuzzy probability range. Some answers indicate ranges instead of one point, like Figure 3.1.k. Therefore they cannot be used for the Bayesian Network. And in some cases it is not obvious what the expert meant to indicate on the scale, like Figure 3.1.f.

3.3.2. Reliability

If an expert indicates his certainty on the scale, the interpreted answer of the elicitor is not the answer to the question that was posed, and therefore not reliable. Besides that, more than half of all questions in the EPIZONE project was answered on an anchor. It is not known why this occurred, but it does raise the question if different anchors on the scale would have resulted in different answers from the experts. The observed cases of two probability questions from the same probability distribution not adding up to one, do also question the reliability of the answer. If the answers do not add up to one, at least one of the answers is not correct.

In the group sessions of the EPIZONE project no information was collected about the process of answering the questions. In the interactive one-on-one setting, the researcher was able to communicate with the expert and was able to correct the expert immediately when noticing a misinterpretation or misunderstanding. When experts are not guided by an elicitor, the experts cannot be corrected anymore.

To conclude, the main problem of the use of the scale in unguided group sessions is the interpretability of the answers from the expert. If the answers cannot be interpreted or if the answers are not reliable, these results are useless for the Bayesian Network.

3.4. Hypotheses

As described in Section 3.3. two main problems are identified. In this section the possible causes of the observations that lead to those problems are hypothesized. If we can determine the possible causes, we can make recommendations to improve the verbal-numerical scale and possibly solve the problems. In the next sections the possible causes of the observation are listed. From these possible causes, some hypotheses are formulated. It appears that the hypotheses can be divided in two

categories: the representation of the scale and the method of the scale. What is meant by the representation and the method is explained in Section 3.4.2. and 3.4.3.

3.4.1. Possible causes underlying the observations

In this section the possible causes of each observation are described, followed by the most likely cause in our opinion.

Frustration

The frustration that was observed in some experts might be caused by the representation of the verbal-numerical scale. For instance, the expert did not understand how to put his answer on the scale, although he had received an instruction about it. The cause of the frustration could also be caused by a lack of expertise of the expert. The questions to provide a point estimate could be the cause of frustration if the expert was not certain enough about the answer to provide a point estimate.

We think the representation of the scale caused confusion by the expert of how to indicate an answer on the scale. This is due to a lack of affordance of the scale; it does not attract intuitive good use of it [14].

Certainty instead of answer

The observation that experts indicate their own certainty instead of the answer could have been caused by the representation of the verbal-numerical scale. The semantics of the verbal labels ‘certain (almost)’ and ‘uncertain’ could have triggered the expert in indicating their certainty on the scale instead of the answer itself. The positioning of the numerical label 100 with the verbal label 'certain' could have caused an association with “I am 100% certain”, that relates to the certainty of the expert about himself.

Furthermore, the available labels on the scale are semantically no good answers to the question. The question is how likely something is to be true. Semantically right answers would be some degree of likely, for instance: very likely, very unlikely, not likely, etc. The numerical labels and the verbal labels like fifty-fifty and uncertain are not semantically good answers to the posed question. This could have caused experts to misinterpret the questions as questions about their certainty.

We think it is most likely that the semantics of the verbal labels ‘certain (almost)’ and ‘uncertain’ caused experts to indicate their own certainty on the scale instead of the answer. It makes the semantics of the scale unclear. This problem does also occur in the Dutch version of the scale, since the Dutch words 'zeker (bijna)' and 'onzeker' have the same meaning.

Wrongly indicated answers

Experts who indicated their answer as an encircled verbal or numerical label could have misinterpreted the relation of the verbal and numerical labels with the vertical line or did not see a relation between them at all. Experts might have treated the verbal labels and the numerical labels as separate scales, instead of integrate the two sets of labels on the vertical line. A possible cause of

people indicating their answer by encircling a verbal label or a numerical label is that an expert does not interpret the verbal-numerical scale as a whole, but rather treats the verbal labels, the vertical line with the horizontal dashes, and the numerical labels each as a separate scales. If so, this is probably caused by the representation of the scale. The verbal labels are positioned relatively far from the vertical line and the horizontal dashes on the vertical line point to the verbal labels but are not at the same height as the verbal labels. This could cause people to interpret the verbal labels as a separate scale. It could also have been that they did not understand how to put their answer on the scale. This again suggests problems with the affordance of the scale; the scale might not attract intuitive good use of it [14].

The wrongly indicated answers could have been caused by experts' misinterpretation of the instructions (to indicate the estimate with a horizontal dash on the vertical line), or they might have just ignored them. The experts who did not indicate their answer as a point estimate might have been not certain enough of their answer to provide a point estimate. Another option is that the experts did not have a point estimate in mind as their answer.

We think the experts themselves did not formulate a point estimate as an answer to the probability questions before they indicated the answers on the scale. Furthermore we think the representation of the scale does not encourage users to use the space between the anchors. We think the positioning of the parts of the scale could cause users to treat the parts of the scale as separate scales of parts instead of treating the scale as a whole.

Answers on anchors

The observation of experts indicating their answers on anchors, could have occurred because the experts really did think that the specific anchor was the answer. It could also have been that the answer they had in mind was close to one of the anchors and they adjusted it towards the anchor, or the expert thought he was only allowed to use the anchors. Another possible cause is that the expert did not know how to indicate another answer besides the available anchors on the scale. They did not exploit the continuity of the scale; the possibility to mark any degree of probability as an answer by putting a dash on the vertical line. This might be caused by the unusual intervals of the numerical labels. The irregular pattern of horizontal dashes on the vertical line might not present the vertical line as a continuous scale, but more as some possible answers on a vertical line. It also might give the impression of the seven numerical labels being somehow special. Arnheim [15] states that we draw on past experience when we use visual perception. What you think an image represents depends on what it reminds you of. For the image to be perceived as a scale, the objects in the image need to be organized as is usual for a scale. Scale anchors are often assumed to be balanced. The expectation is that the labels have equal intervals, where there are opposite poles with a neutral midpoint [16]. The numbers on the scale have unusual intervals for a scale (intervals: 15, 10, 25, 25, 10, 15).

We think the experts adjusted their answer to fit an existing anchor, because they are drawn towards the anchors and in their opinion the anchors are acceptable answers to the questions. By this we do not mean that the experts do not have enough experience or knowledge to form an answer by themselves. We mean that the assessments in this domain leave many factors out. We think there might not be one right point-based answer, but the right answer lies somewhere in a probability

range. Furthermore, we think that the unusual interval of the numerical labels and the irregular pattern of horizontal dashes contributed to experts answering on anchors.

Probabilities not adding up to one

Experts might have not understood that the two answers should have added up to one. This could indicate that the expert might not have understood the concept of probabilities. Another option is that the expert did not check his answers and just overestimated or underestimated his answers. Overestimation is a commonly observed problem with people estimating probabilities [11]. In the verbal-numerical scale the expert is expected to provide a point estimate, by marking the vertical bar. Although the scale includes a list of words to support the expert in giving this mark, it is still wanted that the expert provides a point estimate.

We think the experts did not recognize that the answers, indicated on a verbal label, reflected a numerical probability and therefore they did not calculate with the answers to check whether they added up to one. In most cases people tend to express probabilities verbally; explain them in words rather than in numbers [11]; this was also the motivation for the developers to design a verbal-numerical scale. In order to support the users, the relation between the verbal probabilities and the numerical probabilities should be clear. It is however hard to accomplish that because verbal estimators have subjective numerical values; every verbal estimator has a different meaning for every person in every situation [17]. The researchers were aware of this, and therefore chose to represent the verbal labels not as direct translations of a numerical probability. The verbal labels represent a stable rank ordering. Along with the freedom that experts have, to indicate their answer at some place along the vertical line, the researchers assumed that the fact that verbal estimators have subjective values would not cause problems with the use of the scale.

3.4.2 Hypotheses concerning the representation of the scale

In this section is explained what we mean with the representation of the scale. Then the hypotheses are outlined that are investigated in this thesis. For every hypothesis is described how it is tested in the experiment.

Representation

With the representation of the verbal-numerical scale we mean the placement of all objects of the scale in relation to each other, the interval of the labels, the semantics of the labels and the affordance of the scale. We consider verbal labels, numerical labels, the vertical line and the dashes to be objects. The verbal-numerical scale is designed carefully. The choice of the amount of labels and which labels to use are well investigated and supported. The developers did, however, not realise that the actual visual representation of the scale was that important/would make that much difference. But actually, the representation of the scale determines the way the scale will be perceived by the users [14], and is therefore worthwhile to design carefully.

Hypotheses

Some of the possible causes underlying the observations that lead to the interpretation and reliability problems are investigated in this thesis. In this section we list the hypotheses and describe how they will be tested in the experiment.

Hypothese 1: The affordance of the original verbal-numerical scale does not attract users to indicate their answer as a point-based answer on the vertical line on the scale.

Hypothese 2: The affordance of the original verbal-numerical scale does attract users to indicate their answer on one of the available anchors.

In the experiment we will not instruct the subjects how to indicate their answer on the scale. The method and place of indication of the answers from the subjects will show what kind of use the scale attracts. We ask the subjects to think aloud while they answer the probability questions and while they indicate their answer on the scale. The verbalized thoughts of the subjects hopefully provide information about the reason why subjects indicate their answers as they do. We also designed an alternative version of the scale with a different representation. The scale developed by Renooij and Witteman (Figure 2.3.) from now on will be referred to as the original representation of the scale or the original verbal-numerical scale. In the experiment we will investigate whether the alternative representation of the scale attracts different use of it than the original representation of the scale.

Hypothesis 3: The alternative representation of the verbal-numerical scale results in less wrongly indicated answers and answers indicated on anchors than the original representation.

We think the verbal-numerical scale would result in less answers on anchors if it had equal intervals of the numerical labels and a regular pattern of horizontal dashes. Furthermore we think the verbal-numerical scale would result in less wrongly indicated answers if the relation between the different parts in the scale was more clear. In the next chapter it is described what changes have been made to the representation. It should be noted that we made several changes at once. This means that the effect of the individual chances cannot be measured. With the design of the alternative scale we want to make a possible step towards a solution to the problems, instead of just investigating the possible causes.

Hypothesis 4: The semantics of the verbal labels ‘certain (almost)’ and ‘uncertain’ causes users to indicate how certain they are of themselves on the scale instead of indicating the answer to the question.

This hypothesis will not explicitly be tested in the experiment. By letting the subjects verbalize their thoughts during the experiment, we can detect whether the subjects indicate their certainty instead of the answer itself. In the think aloud we hope to discover whether the reason for this behavior are the semantics of these verbal labels.

3.4.3. Hypotheses concerning the method

With the method we mean providing experts with a scale that combines probability words and numbers as a means to support experts in providing a point estimate. It is about the task of providing a point estimate and the use of the concept of probability in the question and answer.

Hypothesis 5: When asked for a probability assessment, users provide an answer in a format different from a point estimate.

In the experiment the subjects are asked to answer some probability questions and write the answer down on an empty sheet of paper. We will then investigate what type of answers are given by the subjects. If subjects write down a point estimate, they will be asked whether they really mean one point, to ensure that subjects thoughtfully write down a point estimate and really mean a point estimate. The verbalized thoughts hopefully give us insight in the way subjects retrieve information and generate a response before they format it.

Hypothesis 6: The combination of probability words with probability numbers in the verbal-numerical scale causes confusion with the use of the scale.

The subjects are encouraged to think aloud during the whole experiment. The part where they think aloud while indicating their answer on the scale will hopefully show us if the combination of probability words with probability numbers causes confusion or actually helps the subjects to indicate their answers.

Hypothesis 7: Users adjust their answer to fit an existing anchor.

In the experiment the subjects have to write down their answer on an empty sheet first. Later they have to answer the probability question again and indicate the answer on the verbal-numerical scale. By comparing the first (open) answer, with the answer indicated on the scale together with the verbalized thoughts, we can determine if the subjects did adjust their answer to fit an existing anchor on the scale.

Chapter 4: Method Experiment 1

We conducted a qualitative think aloud experiment in order to gain information about the thoughts and motivation of subjects while answering probability questions on the verbal-numerical scale. During the experiment the subjects were instructed to continuously verbalize their thoughts, which is called thinking aloud. The experiment was conducted as is described by Ericsson and Simon [18]. The thinking aloud will hopefully give insight in the cognitive processes of answering questions. The whole experiment was recorded. The think aloud recordings provide insight in the motivation of the subjects' actions and help to determine what causes the observed problems with the verbal-numerical scale. The subjects are experienced veterinarians and in the experiment they answered probability questions in the domain of classical swine fever. We used real experts in this experiment, to stay as close as possible to the original task of experts answering probability questions in their domain of expertise. The subjects answered the questions using two versions of the verbal-numerical scale, the original one and an alternative one we developed for this experiment. This chapter discusses the experimental set-up, design, materials, procedure and participants of the experiment.

4.1. Design

A think aloud probability elicitation experiment was designed in which one independent variable is varied during the study: the representation of the verbal-numerical scale. With the alternative representation we want to make a step towards a possible solution to the problems, instead of just investigate what the causes are of the problems. Furthermore we designed our own realistic probability vignettes for this experiment. This section first describes the design of the alternative representation of the verbal-numerical scale. Secondly, the design of the probability vignettes is outlined.

4.1.1. Designing the alternative verbal-numerical scale

We designed an alternative representation of the verbal-numerical scale, which we think prevents for the problems encountered with the use of the original one. With the representation of the verbal-numerical scale as possible cause of the problems in mind, some changes have been made to the verbal-numerical scale, like the position of the anchors and the number of anchors. The original

and the alternative verbal-numerical scale are depicted in Figure 4.1. The scales with the verbal probabilities in Dutch can be found in Appendix A.

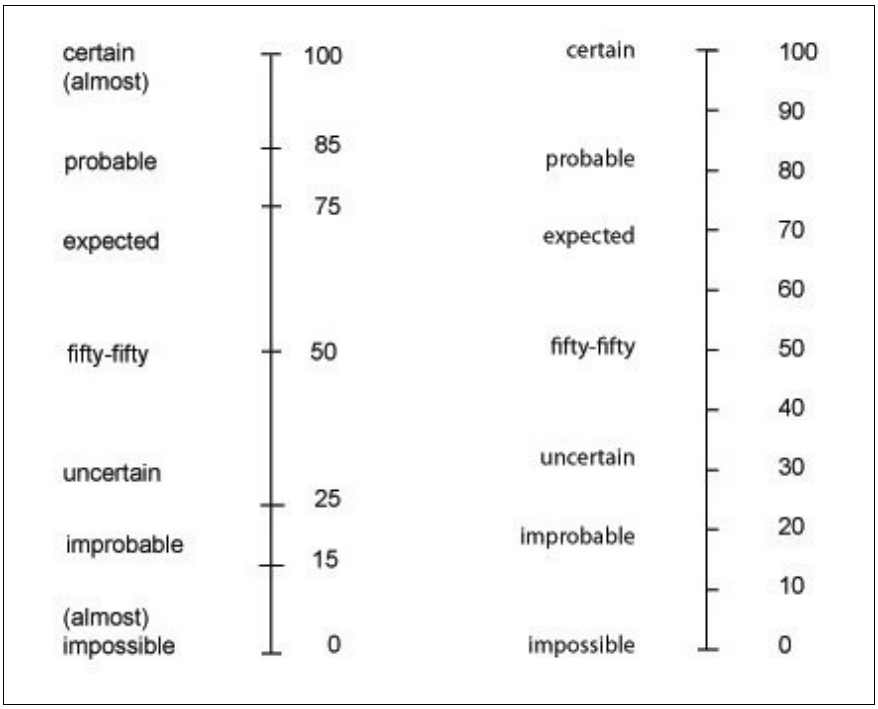


Figure 4.1. The original verbal-numerical scale (left) and the alternative verbal-numerical scale (right).

The changes that have been made in (the representation of) the original verbal-numerical scale are listed below. In the alternative representation:

1. a scale with ten numerical labels is designed with equal intervals, opposite poles and a neutral midpoint;
2. both the verbal labels and the numerical labels are aligned towards the vertical line;
3. the dashes on the vertical line are depicted on the right side of the vertical line, indicating they belong solely to the numerical labels;
4. the vertical line is depicted in the centre between the verbal and the numerical labels;
5. the modifier '(almost)' is removed from the endpoint labels.

The first change is made to give the scale a more common scale-like look. We chose to make the horizontal dashes and correspondent numerical labels at equally distance on the vertical line, with an interval of 10 between each numerical label.

The second, third and fourth change all have to do with making the relation between the different parts of the verbal-numerical scale more clear. We aligned both the numerical labels and the verbal labels towards the vertical line, so the distance between the each label and the vertical line is the same. Furthermore we depicted the horizontal dashes on the vertical line on the right side of the vertical line, to indicate that they solely belong to the numerical labels. The dash belonging to 100 and 0 is depicted as in the original scale, because these numerical labels do relate to the verbal labels 'certain' and 'impossible'. We think this will connect the verbal labels and the numerical labels with each other and with the vertical line.

The fifth change we made in the representation of the scale was removing the modifiers from the verbal labels 'certain (almost)' and '(almost) impossible'. We think that the description of the verbal labels might be confusing for the user of the scale. The modifier 'almost' is enclosed in brackets, which usually indicates that it is not that important or that it is not essential. But as a label, it is important, because it is used to interpret the verbal-numerical scale and to indicate an answer on the scale. It visually causes that the two verbal labels are using two lines, while the other labels only use one.

4.1.2. Designing the probability questions

The subjects were presented with probability questions in their domain of expertise, concerning the diagnosis of classical swine fever.

In collaboration with two veterinarians (who did not participate in the experiment) and a veterinary epidemiologist, familiar with both CSF and Bayesian Networks, we developed vignettes with probability questions to be used in the experiment. The probability questions concern real probabilities from the network which makes the task of assessing those probabilities realistic. The vignettes consist of a description of symptoms of some pig followed by a question about the probability of some other symptom given the described symptoms. The number of symptoms is varied per assessment. To prevent for anchoring biases, the vignettes were designed to have no relation to each other. The questions are each presented on a different sheet of paper. The vignette below is an example of a question involving two symptoms (circulation disorder and cyanosis).

> Example 4.1. probability question
> Imagine a pig with a circulation disorder. How likely is it that this pig shows cyanosis (blue/purple discoloration) of the ears and/or on other body ends (feet, nose, around the tail)?

An overview of the vignettes is added in Appendix B (in Dutch). All subjects were presented with the vignettes in a different order to prevent questions influencing each other.

4.2. Experimental setup

The experiment consistsed of four parts that were preceded by a training session and concluded by an evaluation. The whole experiment was in Dutch and was conducted with Dutch speaking experts. In the training the method of thinking aloud was demonstrated to the subjects and the goal of this method was explained. The subjects were presented with several assignments to practice thinking aloud. See Appendix C for an overview of the assignments of the training part. Below we will briefly explain the four parts of the experiment.

- **Part 1:** the subject was asked to assess 5 (or 6, if he would need the spare question) probabilities that were each presented on a new sheet while thinking aloud. The answer format was left unspecified.

- **Part 2**: the subject had to answer the same questions as in part one, but this time he was forced to indicate his answer to the probability questions on a verbal-numerical scale, without explanation of how to do this.
- **Part 3:** the recordings of Part 1 were played back to the subject per assessment. The subject was confronted with his own thinking aloud and instructed to comment on that.
- **Part 4**: the elicitor translated each answer the subject had indicated on the verbal-numerical scale into a point estimate and explained the subject how his answer was interpreted. The subject was asked to respond to this.

After Part 4 some evaluation questions were posed to the subject to encourage the subject to talk about the scale and the experiment.

By representing half of the subjects with the original verbal-numerical scale and the others with an alternative verbal-numerical scale we aim to find differences in the way subjects respond to the scale.

4.3. Variables

Recall that the independent variable in the experiment is the version of the verbal-numerical scale.
The dependent variables are:

I. Written open answers: the answers written down in Part 1. The answers in the open space could show in what kind of format the subjects answer probability questions. These open answers can later be compared to the answers on the scale to determine whether the subjects adjusted their answers towards anchors.
II. Answers on both versions of the verbal-numerical scale: the answers indicated on the scale in Part 2. The answers on the verbal-numerical scale show what kind of answer the scales attract if the subjects are not instructed beforehand.
III. Think-aloud and all remarks made by the subjects during the experiment. They cover together the 'why' of every action. They provide information about the cognitive processes of answering a question (as described in section 2.3). Furthermore they give insight in the way both versions of the verbal-numerical scale are interpreted. With the think aloud we mean the verbalized thoughts during the experiment. The remarks from the subjects cover all comments of the subjects beside the thinking aloud. They cover for instance the subjects reaction to his thinking aloud in Part 3 and the reactions to the translated answer from the elicitor in Part 4.

4.4. Materials

Several materials have been developed for this experiment, being a protocol, the question-answer sets with additional information, evaluative questions and a consent form.

4.4.1. Protocol

During the experiment a protocol is followed to ensure consistency in all the experiments. The protocol captures the literal phrases the elicitor has to say to the subject. In between also the (possible) actions are described, like handing material to the subject and a list of neutral sentences the elicitor could say when a subject stops thinking aloud. The protocol is added in Appendix D (in Dutch). The directions to the subject in the protocol are directly translated from the directions for thinking aloud experiments described by Ericsson and Simon [18].

4.4.2. Question-answer sets

For every subject a set was composed with six probability vignettes and answer sheets with one of the versions of the verbal-numerical scale. The vignettes in Part 1 are each presented on a different sheet of paper followed by the instruction of the elicitor to provide an answer to the question under consideration, as if it were for a fellow veterinarian. The elicitor, guiding the experiment, had extra information about every vignette to ensure consistent answers if a subject would have questions. The vignettes with the extra information are added in Appendix B (in Dutch).

4.4.3. Evaluation questions and consent form

After Part 4 five evaluation questions were asked to the subjects. They were asked about their opinion concerting:

- The probability questions
- The expression of the open answer
- Indication the open answer on the scale
- Combination of words and numbers on the scale
- The experiment

The subjects in addition, were asked to sign for permission of use of the recorded and written material for the purpose of supporting the experimental results. This consent form is added in Appendix E (in Dutch).

4.5. Setting

The experiment is conducted in a one-to-one setting with the elicitor and the subject. The camera (a Sony Digital 8 DCR-TRV345E) is positioned behind the subject on a tripod in such a way that can be recorded what the subject is writing down. It is mentioned to the subject that the video recorder is only directed at the hands of the subject, to capture the actions on paper. The experiment is also recorded by a stereo microphone (Sony), and a voice recorder (Olympus VN-3100PC). The voice recorder is able to play and record and is therefore used for playing back recordings to the subject. The microphone on the table is connected to the video camera and takes care of an excellent stereo recording of both the elicitor and the subject synchronously with the video. A triplex board is placed in front of the subject. Strokes of foam board on the right and top of the triplex board make sure the sheets the subjects works on during the experiment stay within the region the camera can record. The setting is shown in Figure 4.2.

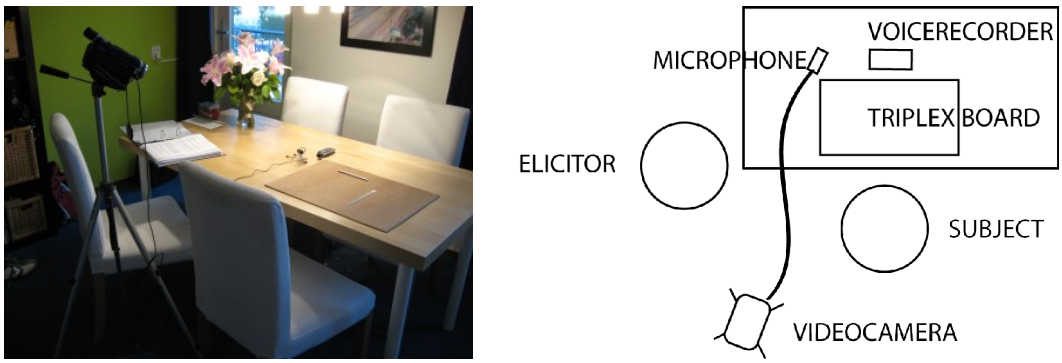


Figure 4.2. Experimental setting.

4.6. Procedure

The elicitor introduces the experiment by stating that from that moment on everything will be read from a paper protocol to ensure consistency between the experiments. The protocol is added in Appendix D (in Dutch). Then the subject is told that the experiment is anonymous. Next, it is explained what thinking aloud means and what the task of the subject is. Before the experiment is started it is outlined that our concern is the process of answering the questions and how these answers are put on paper. It is clearly stated that the answer itself is not our interest and that a questions cannot be wrongly answered. Every answer is correct. The subject is, however, encouraged to do his best to answer the questions as good as possible.

Some exercises are executed to practice the thinking aloud. In order to test the recording devices and the subject getting used to being recorded, these practices are recorded. In Part 1, the subjects are instructed to flip the paper after writing the answer down and go on with the next question. At that stage they are not allowed to look back at assessments they have already made. In Part 2 the subjects are instructed to answer the questions from part one again. Only this time they have to answer on the scale. The subjects are in this part presented with the questions, their open answers and the scale. Every question is indicated on a new sheet with the scale. The subjects are again encouraged to think aloud while executing Part 2. In Part 3 the subjects listen to the recordings of

their verbalized thoughts from Part 1. The recording is paused after every question and the subjects are than asked to tell everything they remember about giving their answer. In Part 4 the elicitor shows the subjects how their indicated answers on the scale are interpreted. The elicitor measures the distance from '0' to the indicated answer with a ruler, regardless of the method the answer was indicated. The subjects are asked to comment on this interpretation. This comment is the starting point to a conversation about the representation and use of the scale. To conclude, some evaluation questions are asked about the experiment and the use of the scale.

After the experiment all subjects were surprised with a small present.

4.7. Subjects

Six experienced veterinarians were willing to participate in the experiment. Below a summary is made of the subjects experience and some demo graphical information. All subjects are male. Five subjects are aged over fifty, one was between forty and fifty years old. Five subjects had never done this kind of assessment, one subject said to have had some experience with estimating probabilities. None of the subjects ever worked with the verbal-numerical scale. On average, the subjects have 26 years of practice as a veterinarian, the subject with the least experience has 15 years of experience, and the most experienced subject has 37 years of experience. They all have at least experienced one outbreak of classical swine fever. The number of experienced outbreaks is between one and eight. An overview of the subjects is found in Table 4.1.

	Scale	Gender	Age	Years of experience	Experience prob.quest.	Experience with scale	Number of outbreaks
1	Alternative	Male	> 50	25	No	No	8
2	Original	Male	> 50	37	Yes	No	4
3	Alternative	Male	40-50	21	No	No	2
4	Original	Male	> 50	32	No	No	3-4
5	Alternative	Male	> 50	15	No	No	1
6	Original	Male	> 50	30	No	No	5

Table 4.1. Overview of the subjects (1-6), which representation of the scale was presented to them (alternative, original) and their experience.

Chapter 5: Results Experiment 1

This chapter presents the results of Experiment 1. As a reminder we first briefly repeat the general structure of the experimental design (Part 1-4) and it's output. Second, the structure of this chapter is outlined.

The structure of the experimental design is as follows:

- In Part 1 subjects provided written open answers to the probability questions while they were thinking aloud.
- In Part 2 subjects read the probability question again and answered on the scale while they were thinking aloud.
- In Part 3 the subjects listened and commented on their own thinking aloud during Part 1.
- In Part 4 the elicitor translates the answer of the subject to a point estimate. The subject is asked to react which leads to a discussion of the verbal-numerical scale.

To structure the results in the format of utterances from the think aloud, these results are described on the basis of the cognitive processes earlier discussed in Section 2.3. Figure 5.1 depicts the cognitive processes that took place during Part 1 and Part 2 of the experiment.

The structure of this chapter is as follows. Section 5.1. describes and analyzes the output (written open answers) and the utterances from Part 1 and 3 that relate to the cognitive processes that took place in Part 1. The cognitive steps include interpreting the question, retrieving information and generating the answer.

Section 5.2. describes and analyzes the output (answers on the scales) and the utterances from Part 2 and 4 that relate to the cognitive processes that took place in Part 2. The cognitive steps include interpreting the question, retrieving information, generating the answer, formatting the answer and indicating the answer.

Finally, Section 5.3. describes the results of the evaluation at the end of the experiment.

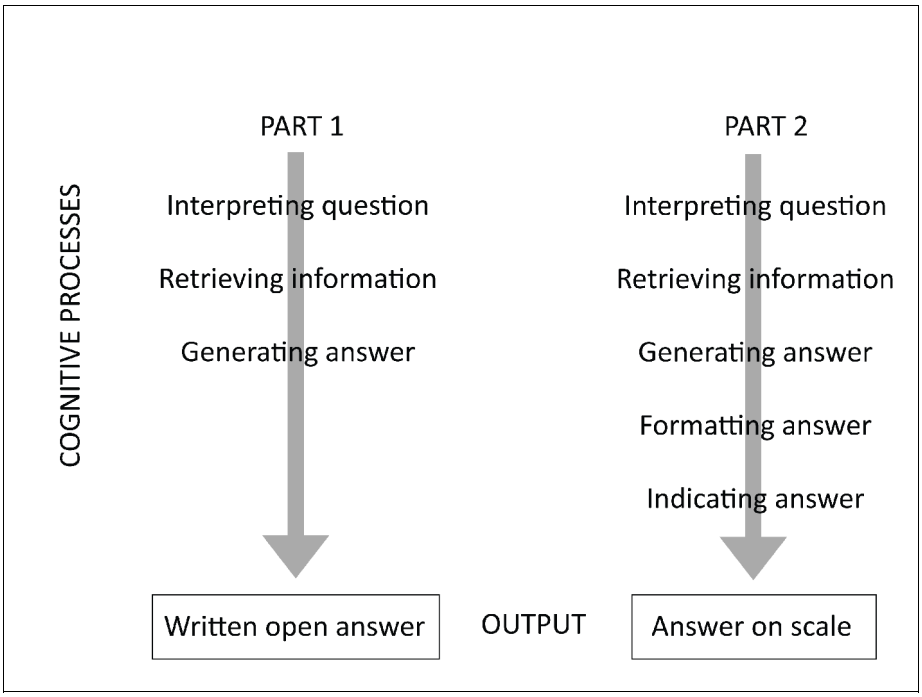


Figure 5.1. Graphical representation of the cognitive processes and the output from Part 1 and 2 of the experiment.

5.1. Written open answers

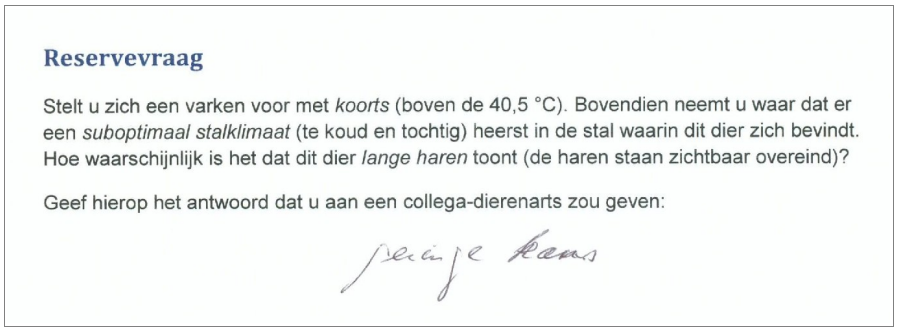
In 5.1.1 we first analyze the open answers that are written down in Part 1. Before the open answer was written down, the subjects had to interpret the question, retrieve information and generate the answer. These steps can be traced by analyzing the think-aloud from Part 1 and 3. We describe how the subjects walked through these steps in Section 5.1.2.

5.1.1. Analysis of the written open answers

Four answer types are distinguished, examples of which are shown in Figures 5.2-5.5 (in Dutch):

1. N - numerical answers (Figure 5.2.)
2. V - verbal answers (Figure 5.3.)
3. NR - numerical-range answers (Figure 5.4.)
4. V + N - verbal + numerical answers (Figure 5.5.)

Figure 5.2. Example of a numerical answer (N)



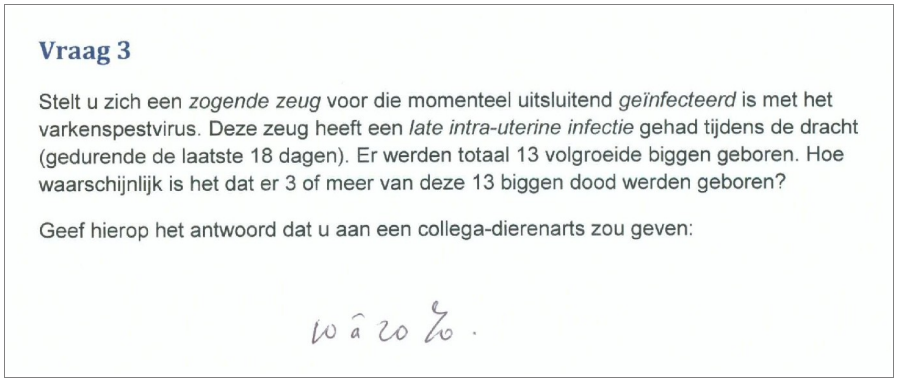
Reservevraag

Stelt u zich een varken voor met *koorts* (boven de 40,5 °C). Bovendien neemt u waar dat er een *suboptimaal stalklimaat* (te koud en tochtig) heerst in de stal waarin dit dier zich bevindt. Hoe waarschijnlijk is het dat dit dier *lange haren* toont (de haren staan zichtbaar overeind)?

Geef hierop het antwoord dat u aan een collega-dierenarts zou geven:

zeer grote kans

Figure 5.3. Example of a verbal answer (V)



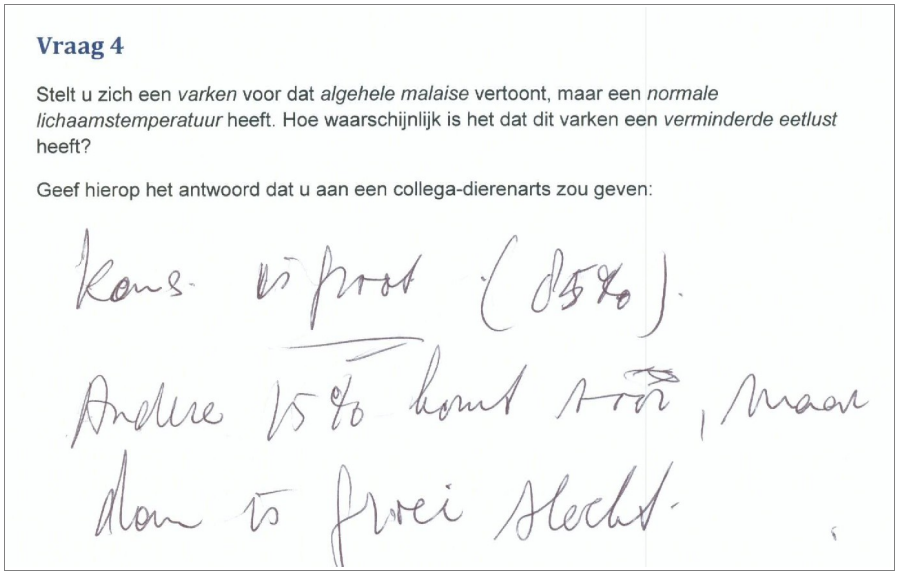
Vraag 3

Stelt u zich een *zogende zeug* voor die momenteel uitsluitend *geïnfecteerd* is met het varkenspestvirus. Deze zeug heeft een *late intra-uterine infectie* gehad tijdens de dracht (gedurende de laatste 18 dagen). Er werden totaal 13 volgroeide biggen geboren. Hoe waarschijnlijk is het dat er 3 of meer van deze 13 biggen dood werden geboren?

Geef hierop het antwoord dat u aan een collega-dierenarts zou geven:

10 à 20 %.

Figure 5.4. Example of a numerical range (NR)



Vraag 4

Stelt u zich een *varken* voor dat *algehele malaise* vertoont, maar een *normale lichaamstemperatuur* heeft. Hoe waarschijnlijk is het dat dit varken een *verminderde eetlust* heeft?

Geef hierop het antwoord dat u aan een collega-dierenarts zou geven:

Kans is groot (85%).
Andere 15% komt ook voor, maar
dan is geval slecht.

Figure 5.5. Example of a verbal + numerical answer (V + N)

Table 5.1. shows the answer type of each question per subject. The results show a clear preference per subjects for a certain answer type. Three subjects expressed all questions in the same answer type. Subject 2 (S2) used only numerical ranges (NR) as answer type. Subject 3 (S3) answered all questions with a number (N) and Subject 4 (S4) answered all questions verbally. Subject 1 (S1) and Subject 5 (S5) provided varying answer types. Subject 6 (S6) mainly used words in his answers (V), and in two cases he added a number to it (V + N). From this we conclude that there is no correlation between the type of answer and the question.

	S1	S2	S3	S4	S5	S6
Q1	NR	NR	N	V	V	V
Q2	NR	NR	N	V	NR	V + N
Q3	NR	NR	N	V	NR	V + N
Q4	NR	NR	N	V	V	V
Q5	NR	NR	N	V	NR	V
Spare	V	NR	-	V	-	V

Table 5.1. Overview of the answer type the subjects (S1 – S6) gave to the experiment questions (Q1 - Q5) and the Spare question: Numerical (N), Verbal (V), Numerical Range (NR) or Verbal and Numerical (V + N). Subjects 3 and 5 did not answer the Spare question.

Table 5.2 shows the frequency of the four answer types. Five open answers were expressed numerically, by providing a numerical point estimate. Thirteen answers were expressed verbally, ranging from a single term like 'small chance' to a page full of comments. Fourteen answers indicated a numerical range, like '< 5%' or '20-30%'. Twice a subject provided a verbal answer along with a numerical answer.

Answer type	# Answers
Numerical (N)	5
Verbal (V)	13
Numerical range (NR)	14
Verbal + numerical (V+N)	2

Table 5.2. Frequency of each answer type.

5.1.2. Cognitive processes

The think-aloud from Part 1 and 3 provides information about the process of answering the probability questions. By letting the subjects think aloud in this experiment we tried to capture this process. In this section we describe how the subjects interpreted the question, retrieved information and generated their open answer. Six probability questions (1-5 and spare question) are in a total of 34 times answered. The tables in this Section show the occurrence of some action and the amount of answered questions.

Interpreting the question

In order to answer a question, the question needs to be interpreted in order determine what is asked. From analyzing the think-aloud of Part 1 and 3 we looked for utterances that would reveal that the subject interprets the question. Table 5.2. shows five categories of utterances that reveal interpretation of the question and the amount of times they occurred with a question. Questions can be interpreted by utterances with actions from more than one category.

Category	# Occurrence
Define variables that are used in the question	3
Debate the used terminology	6
Adjust the question	2
Repeat (parts of) the question	12
Check interpretation with the elicitor	1

Table 5.2. The times of occurrence of the categories of utterances made by the subjects in the think-aloud that reveal interpreting the question.

We will now give an example in each category of phrases uttered by subjects (translated from Dutch transcriptions) that reveal the taken action.

In Table 5.2. it is shown that with 15 question no utterances were found in the think aloud that revealed the subject interpreted the question. In these 15 cases the subjects immediately started with answering the question. Example 5.1. shows a quote from the think-aloud that shows the subjects starts with answering the question right after reading it.

Example 5.1. No retrievable interpretation
Subject reads the question out loud and says: "*Well, that chance is very high because the pig is cold and has a fever so I estimate that to be a chance of about 100% almost.*"
(Subject 3)

Three times a subject defined the variables in the question to himself. An utterance of this category can be read in Example 5.2.

Example 5.2. Define variables that are used in the question
"*overall malaise, that means the animal has grown lean, has no good skin color, has thicker hair."* (Subject 4)

Six times the terminology that was used in the question was debated. The terminology of Question 3 (about reduced appetite) and that of the Spare Question (about long hair) were debated the most. Example 5.3. contains an utterance illustrating a debate of the used terminology.

Example 5.3. Debate the used terminology
"*That is...the terminology is a little unusual. An overall malaise is very general, in the terminology it is always an animal that shows abnormal behavior, so the probability that it has a declined appetite is reasonably big. Unless you formulate the idea of malaise better*".
(Subject 2)

Two times the question was adjusted by subjects. One time it was adjusted by leaving a variable out and one time by dividing the question into two questions. In Example 5.4. Subject 4 adjusts the

question by leaving the variable 'conjunctivitis' out. He gives an answer concerning the relation between an infection of the mucous membrane and tear stripes while he was asked for the relation between an infection of the mucous membrane and a visibility of conjunctivitis through tear stripes.

Example 5.4. Adjust the question
"(...) an infection of the mucous membrane of the upper respiratory. The conjunctivitis has no relation with that. (…) So it has an infection of the mucous membrane, then I say because of that infection there is a discharge of tears. It will produce abundant lachrymal fluid that is insufficiently discharged. Thus the chance it has stripes from tears is very high."
(Subject 4)

Twelve times a subject repeated parts of the question after reading it. Example 5.5. shows an example of a subject repeating parts of the question.

Example 5.5. Repeat (parts of) the question
"*Give the answer you would give to a fellow-veterinarian. With a mucous of the upper respiratory ...how likely is it that this pig shows conjunctivitis.*"
(Subject 2)

One time a subject asked the elicitor if had interpreted the question right. Interpreting the question also contains interpreting what kind of answer is asked. Example 5.6. shows that Subject 1 asks if a numerical answer is expected from him.

Example 5.6. Check interpretation with the elicitor
"*How likely is it that the animal has long hair. Should that be expressed as a number?*"
(Subject 1)

Information retrieval

To answer a question, subjects have to retrieve information from their memory to be able the generate an answer. From analyzing the think-aloud of Part 1 and 3 we looked for utterances that would reveal that the subject retrieves information. Table 5.3. shows two categories of utterances that reveal retrieving information and the amount of times they occurred with a question. Information is retrieved with actions from only one category.

Category	# Occurrence
Reasoning about the variables in an abstract situation	22
Reasoning by recalling concrete situations	4

Table 5.3. The times of occurrence of the categories of utterances made by the subjects in the think-aloud that reveal retrieving information.

We will now give an example in each category of phrases uttered by subjects (translated from Dutch transcriptions) that reveal the taken action.

In 8 cases it could not be observed and not categorized that the subject retrieved information from his memory to answer the question. The subjects immediately started with formulating their answer to the question, as is shown in Example 5.7.

Example 5.7. No observable information retrieval
"(...) provide the answer you would give to a fellow veterinarian. I estimate that chance thus reasonably big, the last eighteen days. So I estimate that three of the, well at least one, I estimate about eighty, eighty percent."
(Subject 3)

For 22 cases the information retrieval consisted of what-if reasoning about the variables from the questions. While reasoning the pig under consideration stayed an abstract pig, like is shown in Example 5.8. For 4 cases concrete situations where recalled from memory as information retrieval. Example 5.9 shows an example.

Example 5.8. Reasoning about the variables in an abstract situation
"In my logic as I look at it now, when you have an infection there in the last days of the pregnancy, eighteen days, and mature piglets are born, then it is rather unlikely that more than three of them are dead, because the infection of the swine fever has had no influence on that. According to my knowledge early infections lead to death, but late ones do not."
(Subject 5)

Example 5.9. Reasoning by recalling concrete situations
"I reminded from my own past from swine fever."
"(...) because I stand relatively often in the abattoir. Then you sometimes see pigs with terrible lungs and heart defects and then you think how is it possible that this pig walked in here, as a matter of speech. Because you already saw them in the stables."
(Subject 3)

Generation of the answer

To answer a question, subjects have to generate an answer. From analyzing the think-aloud of Part 1 and 3 we looked for utterances that would reveal the way the subject generates his answer. Table 5.4. shows three categories of utterances that reveal a way to generate an answer and the amount of times they occurred with a question. Utterances that show the generation of an answer fall for each question in just one category.

Category	# Occurrence
Generate the answer by some kind of calculation	3
Generate the answer by imagining a real life situation	1
Generate the answer in relation to another estimate	1

Table 5.4. The times of occurrence of the categories of utterances made by the subjects in the think-aloud that reveal generating the answer.

We will now give an example in each category of phrases uttered by subjects (translated from the Dutch transcriptions) that reveal the taken action.

The generation of the answers was for most questions not retrievable from the think-aloud. In 30 cases the subjects provided an answer with no preceding action that indicated how this answer was generated. With 3 questions with 3 different subjects the answer was calculated, like illustrated in Example 5.10. In only one case a subject based his answer on an estimate of a related situation, see Example 5.11. In one case a subject generated his answer on the basis of imagining a real life situation, like in Example 5.12. One time subject did more than one action during the generation of the answer for one question. When looking closely at the think aloud of the subject we noticed that the subjects expressed the probability that was asked for very often as some size of a chance, like 'chance is big' (Dutch: kans is groot). Apparently subjects translate the probabilityquestion immediately to a question about chances and further express the answers in the size of that chance.

Example 5.10. Generate the answer by some kind of calculation
"*Well, ten percent, that will be one point three. Therefore it is, the chance of three dead ones is, that is cumulative, so that chance is obviously much less than ten percent.*"
(Subject 5)

Example 5.11. Generate the answer by imagining real life situation
"At that moment I just imagine a brood in a similar setting in a stable. Some experience is then added. I can imagine it"
(Subject 2)

Example 5.12. Generate the answer in relation to another estimate
"*How likely is it that three or more of these thirteen pigs are born dead. Well that chance is rather large I think, relatively large, because that chance is normally already large (…). The percentage is based on the normal presence of a dead pig in a brood of thirteen.*"
(Subject 1)

5.2. Answers on the verbal-numerical scale

In Part 2 of the experiment the subjects are instructed to answer the probability questions again. This time, they have to put their answer on the verbal-numerical scale. We wanted to investigate how they would do that without explicit instructions. Before the subjects indicated their answer on the scale, the subjects had to walk though some cognitive processes. These steps can be traced by analyzing the think-aloud from Part 2, 3 and 4. We describe how the subjects walked through these steps. In 5.2.1 we first analyze the answers on the scale that were indicated in Part 2. Next, the results about the cognitive process of answering the question are described in 5.2.2.

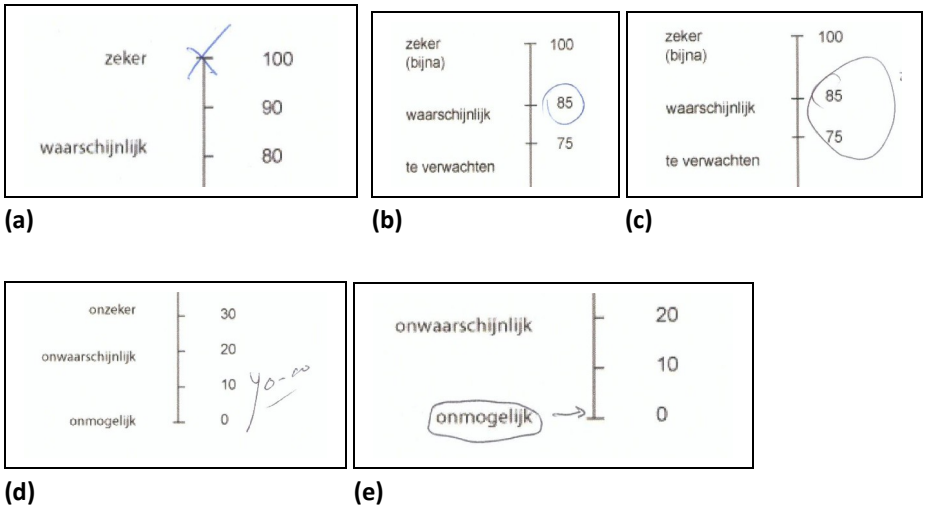
5.2.1. Analysis of the answers on the verbal-numerical scale

This section outlines the answers that were indicated on the verbal-numerical scales. First the methods of indicating the answers are discussed, followed by the type of the answers. For both the method and the type, the differences between the answers on the original and the alternative scale are discussed.

Method of indication

The method of indication refers to how the subjects indicated their answer on the verbal-numerical scale. We found five methods of indicating the answers, examples of which are shown in Figure 5.2.a-e. These five methods are:

1. putting a cross on the vertical line (a)
2. drawing a circle around a numerical label (b)
3. drawing a circle around combinations of labels, dashes and/or the vertical line (c)
4. indicating a range (d)
5. drawing a circle around a verbal label and connect it with a line to the vertical line (e)



(a) **(b)** **(c)**

(d) **(e)**

Figure 5.2. Details of answers on the scale from subjects from the Think Aloud Experiment.

Table 5.5. shows the method for every question per subject and Table 5.6. shows the frequency of every method.

	S1	S2	S3	S4	S5	S6
Q1	4	3	1	2	5	2
Q2	4	2	1	2	4	2
Q3	2	2	1	2	5	2
Q4	4	3	1	3	5	2
Q5	2	3	1	2	5	2
Spare	2	3	-	-	-	2

Table 5.5. The method of indication (1 – 5) for each question (Q1-Q6, Spare question) by each subject (S1-S6). The white colored subject numbers in the gray columns relate to the subjects who were presented with the original verbal-numerical scale.

Method	# Answers
1. Cross vertical line	5
2. Circle numerical label	15
3. Circle combination of labels, dashes and/or vert. line	5
4. Range	4
5. Circle verbal label connected to vert. line	4

Table 5.6. Frequency of each method of indication.

It should be noted that Subject 3 had two attempts for Part 2. In the first attempt it turned out that he filled in his certainty of his open answers. The whole Part 2 was explained again and repeated. Only the answers for the second attempt are used here. Furthermore it should be noted that all subjects except for Subject 6 started Part 2 by indicating how certain they were of their written open answer. After a complementary instruction, the subjects understood the assignment of Part 2.

Most answers were indicated with a circle around a numerical label (15). The other four methods of answering were all used four or five times. Two subjects indicated all answers in the same way. The other subjects used two ways to indicate their answer. No correlation can be seen between the method of indication and the questions.

Scale representation and method of indication

The subjects with a white subject number in the gray columns were presented with the original representation of the scale, while those with a black subject number in the white columns were presented with the alternative representation. The subjects who were presented with the original representation (subjects 2, 4 and 6) only drew circles to indicate their answers (methods 2 and 3). The subjects who used the alternative representation (subjects 1, 3 and 5) mostly indicated point estimates on the vertical line (methods 1 and 5), except for Subject 1 who provided ranges and drew circles around numerical labels.

Type of answer

Method 1 and 5 and are considered point estimates on the vertical line. Method 2 is considered as an answer on an anchor and manner 4 is considered a range. For method 3 it is assessed per answer if it is an anchor or not. In one case it is not obviously an answer on an anchor. This answer is presented

in the Table 5.5. in the category 'other'. The table shows that more than half of all answers were indicated on an anchor. Nine answers were indicated as a point estimate.

	Original	Alternative	Total	Total %
Anchor	16	3	19	57,6%
Point	0	9	9	27,3%
Range	0	4	4	12,1%
Other	1	0	1	3,0%
Total	**17**	**16**	**33**	**100,0%**

Table 5.5. The number of answers that were indicated on an anchor, a point, a range or in another way.

Scale representation and type of answer

Subjects who were presented with the original scale all indicated their answer as an anchor, except for the one answer in the category 'other'. The subjects using the alternative representation together indicated 3 answers as an anchor. Point estimates and ranges were only indicated on the scale by subjects who used the alternative verbal-numerical scale.

5.2.2. Cognitive processes

In Part 2 the subjects have to answer the questions again, but this time they are forced to answer in the format of the verbal-numerical scale. Because they already answered the questions in Part 1, the first three cognitive processes (interpreting the question, information retrieval and generating the answer) are not walked through by most subjects. The results about these three processes are combined in one paragraph. The last two cognitive processes of answering a question are intertwined. These two processes together describe how the subjects formatted their written open answer from Part 1 towards an answer that they indicate on the verbal-numerical scale. To describe this process the written open answers (from Part 1) and the corresponding answers on the scale (from Part 2) were compared and the differences are explained by the think-aloud (from all Parts).

Interpreting the question, retrieving information and generation of the answer

Only two subjects showed that they were interpreting the question again in Part 2. Subject 4 reasoned about the interpretation of one question. Subject 6 debated the used terminology and adjusted 4 questions. For instance the fragment in Example 5.13 shows that the subject adjusts the question about infection of the mucous to a question about a severe infection of the mucous and excludes mild infections that can be observed by hearing the pigs cough. Two subjects showed again what-if reasoning with the variables. Subject 4 reasoned again about the variables with one question because he could not indicate his original written open answer on the scale. In order to decide what numerical answer he should circle he reasoned again about the question. Subject 6 reasoned about the variables of every question again, like he did in Part 1. The generation of the answer is not retrievable from the think-aloud for any subject.

Example 5.13 Adjusting questions

(…) still infection of the mucous membranes of the frontal airways very light cases and very severe cases exist, but if it is a clear infection so not just a superficial light chronic one than the mucous membranes of the eyes will certainly participate, than he has a clear conjunctivitis. If you say I observe kind of a dry coughing in my stables , just a little cough (…) than it is more in the lungs and than the eyes do not have to participate necessarily.
(subject 6)

Formatting the answer and indicating the answer

In Appendix F for every subject the open answer and a description of the corresponding answer on the scale are listed. Our interest is the difference between the two answers. Both are answers to the same probability question. If the answers differ, we want to determine the reason why they differ in order to find out if the open answers were adjusted under influence of the scale. More specifically, we hypothesize that subjects will adjust their open answers to the available anchors on the scale. In many cases, the open answer is a verbal description, while the answer on the scale might be a range or a circled numerical label. Therefore we cannot measure if and to what extent the answer is influenced or changed by the scale, beside from its format. The think aloud provides some reasons for the differences between the answers. Below we will describe per subject the cognitive processes and our assessment of whether the answers were adjusted and why they were adjusted.

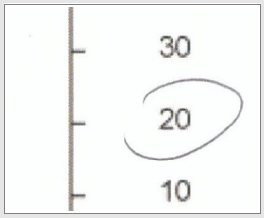
Subject 1 (alternative scale)

Subject 1 wrote down rather small probabilities as open answers, like <0,005% and <0,01%. The answers on the alternative scale in Part 2 were indicated on the side of the numerical labels by circling a number or by drawing a curly brace capturing two numerical labels. The answers he indicated on the scale were in all cases higher than his open answers. For instance, he circled the numerical label '20' as his answer on the scale to the same question he earlier answered with “5-10%” (see Example 5.2.a.) as his open answer. During Part 4 of the experiment it became clear he focused on the verbal labels on the scale. His answers on the scale correspond to the verbal term at the same height. So, the circled numerical label '20' is actually his way of indicating that the probability is 'unlikely'. The answers as interpreted by the elicitor in Part 4 are all incorrect interpretations according to the subject. He considers his open answers however as the correct answers. Subject 1 did not adjust his open answers under influence of the scale. He formatted his open answer as a verbal probability but indicated his answer on the side of the numerical probabilities. From the think-aloud it is clear that he had trouble indicating his answer on the scale because according to him it was hard to interpret the scale.

Example 5.2.a

Open answer:
“5% - 10%”

Answer on scale:



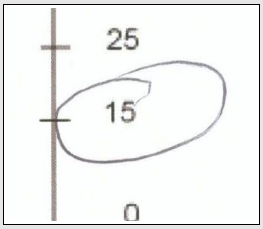
Subject 2 (original scale)

Subject 2 provided numerical ranges as open answers. On the scale in Part 2 he circled numerical labels. His answers on the scale reflect the median of the range in the written open answer or one of the numbers mentioned in the range. For instance, he circled the numerical label '15' on the scale where his open answer was "10-20%" (see Example 5.14.). When the median of his range is not available as a numerical label he would circle the two numerical labels that are directly above and below it. So even when the correct numerical label was not available, he would circle another numerical label, instead of indicating it as a point on the vertical line. Subject 2 agreed on the answers as they were interpreted by the elicitor in Part 4. From the think-aloud it is clear that he generates his answer as the median of his written open answer to fit an existing anchor numerical answer. He adjusted his answers towards the numerical anchors of the scale. Subject 2 had trouble indicating his answer, because he had trouble interpreting the scale. The use of both words and numbers confused him, as is clear from the think aloud fragment in Example 5.14.

Example 5.14

Open answer:
"10-20%"

Answer on scale:



Fragment think aloud:
Subject: "There is something I do not understand"
Elicitor: "*What is it that you do not understand?*"
Subject: "*Look, on the scale a correlation is displayed between unlikely and a number. I have this number in my head and I want to put it somewhere on the scale and then I end up at the crossing with unlikely. While I want to go...that I point that ten and twenty as a probability (…) I think this twist between numbers, yes...*"

Subject 3 (alternative scale)

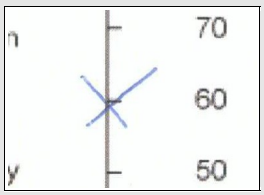
Subject 3 provided point estimates both as written open answer and as answer on the scale. The difference between his open answers and the corresponding answers on the scale are quite large. For instance, the question he earlier answered with '80%' is answered on the scale in the following way: a cross on the horizontal line next to the number '60' (see Example 5.2.d.). The think aloud provides information about why the open answers differ so much from his corresponding answers on the scale. It appears that Subject 3 was more capable of giving a good answer after revisiting the questions, because he had more time to think about the question. He did not adjust his open answer to the scale. He just adjusted his answer to the question. Only the answers on the scale are considered to be correct. The open answers are no longer relevant and replaced by new answers on the scale.

Example 5.2.d.

Open answer:

"80%"

Answer on scale:



Subject 4 (original scale)

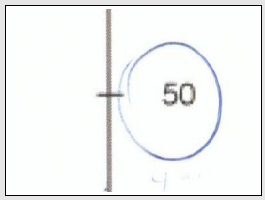
Subject 4 provided verbal written open answers and indicated his answers on the scale by circling numerical labels (see Example 5.15). The answer formats are hard to compare, because no numerical value is mentioned in the open answers. While indicating his answer in some cases he added that he rather would have circled another numerical label if that label would have existed (see Example 5.15). When the think-aloud is read closely it becomes clear the subject thought the existing numerical values were the only answers he could choose. With this knowledge we can conclude that he adjusted all his open answers towards the existing numerical labels. Subject 4 thought it was not possible to indicate his answer between the numerical labels. He was afraid his answer could not be interpreted correctly. In the experiment it was really clear he preferred an answer other than the available numerical labels. Yet he circled an existing numerical label.

Example 5.15.

Open answer:

"*Only when we talk about severe heart failure and/or lung condition, a cyanosis will become visible. This correlation does not necessarily exist.*"

Answer on scale:



Fragment of think-aloud:

"If the number forty would have been written here [subject points between numerical labels 50 and 25], I would have said forty."

Subject 5 (alternative scale)

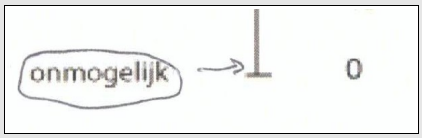
Subject 5 provided numerical ranges and verbal answers as open answer. On the scale he mostly circled a verbal probability label and drew from that circle a line to a point on the vertical line (see Example 5.2.g). In one case he drew some kind of brace connecting two numerical labels. His answers on the scale reflected very precisely his open answer. For instance, his open answer of <1% was indicated by circling the verbal label 'almost impossible' with a line towards a place on the vertical line about two millimeters above the horizontal line indicating the 0%. He said he is visually orientated and likes to draw things. This might be a reason to indicate the answer so creatively. On the other hand he said afterward that it would have been better to delete the words on the scale and only use the numbers. This subject is comfortable using the words but feels the urge to annotate that word with a specific numerical probability.

Example 5.2.g.

Open answer:

"*1% chance*"

Answer on scale:



Subject 6 (original scale)

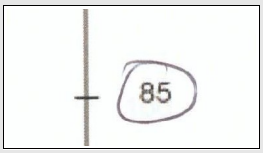
Subject 6 provided verbal open answers, in some cases combined with a numerical answer. On the scale he indicated all answers by circling numerical labels. When comparing the open answers with the answers on the scale some differences were noticed. For instance, Subject 6 wrote down as an open answer 'chance is large (70%)' and circled the numerical label '85' on the scale (see Example 5.2.h). From the think-aloud it was derived that the subject wanted to indicate '70' on the scale, but thought that he only could indicate his answer on the available numerical labels. He did accept both '70' and '85' as good answers to the probability question. Subject 6 thought that it could not be the intention of the scale that the user would indicate answers that are not already listed on the scale. He perceived the numerical labels as the only possible answers (see Example 5.2.i.). He circled the numerical label that is closest to the answer he had in mind.

Example 5.16

Open answer:

"chance is large (70%). Depends on a combination of factors: NH3 + PRRS + influenza + occupation + dust. CSF is only possible with more specific symptoms (petechiae, huddling, high T, anorexia everywhere or specific)."

Answer on scale:



Example 5.2.i.

"*I did not think I was allowed to add a number that is not already there. That is why I chose the higher seventy five percent. If seventy five had not been there and a possibility to indicate seventy would have, then I would have circled seventy*."

Elicitor: "*Why did you think you had to circle a number to indicate your answer?*"

Subject 6: "*There is no other way to express my choice, so I drew a circle. I also thought not to fill in another number, because I assumed these [the labels] are answers I have to choose from."*

To conclude, four categories are observed that describe the way the subjects format the written open answer from Part 1 towards an answer indicated on the verbal-numerical scale.

1. Indicating the written open answer on the scale, but misinterpreting the scale (S1)
2. Adjusting the written open answers towards anchors on the scale (S2, S4, S6)
3. Correcting the written open answer because of new insights or more time to think about the probability question (S3)
4. Conscientiously indicating the written open answer on the verbal-numerical scale (S5).

Subject 1 had looked at the verbal labels of the scale and was not aware that he also indicated percentages on the scale. He misinterpreted the verbal-numerical scale. Subjects 4 and 6 also adjusted their written open answers towards anchors on the scale, like Subject 2. For Subject 4 and 6 however it is clear that they thought the numerical labels were the only possible answers to indicate

on the scale. Subject 3 adjusted his open answers because he had more time to think about the questions and came op with a better answer the second time around. Subject 5 indicated his answer very carefully by encircling a verbal label and connect it with an arrow to a point on the vertical line.

5.3. Evaluational questions

At the end of the experiment some questions were posed to the subject to encourage the subject to talk about the scale and the experiment. The information that was gathered is summarized in this section.

5.3.1. Answers to the questions

The subjects were asked about their opinion of:

- The probability questions
- The expression of the open answer
- Indication the open answer on the scale
- Combination of words and numbers on the scale

The results for those five topics are summarized in this paragraph per topic.

Probability questions

Two subjects thought the questions were clear. Subject 1 thought the questions were unclear, not specific enough, while Subject 2 said it was hard for him to determine what was meant by the used terminology. Subjects 3 and 4 indicated that they thought the questions were clear. Subject 5 had some troubles with the layout of the questions because of the italic words and the additional explanations between parentheses, because of his dyslexia. Subject 6 mentioned that the terminology was not always clear. He thought maybe he was being framed to look how accurately he would answer, or he thought the person who designed the questions had little experience with the terminology. Subject 6 added that he had to read the question carefully to understand what was being asked.

Providing an open answer

Subject 1 mentioned that he was inclined to reason with the diagnoses of classical swine fever in mind. Subject 2 thought that thinking aloud was a good idea, because the elicitor gets more accurate information from the subject. Subject 3 thought it was helpful and fun to provide the open answer first and then the answer on the scale. Subject 4 stated that he as a veterinarian was still used to answer such questions. He explains that if all symptoms of a certain disease are present it is still not certain you are dealing with that particular disease. Classical swine fever is a clear example of such a disease. The symptoms can easily belong to other diseases. You can never be a hundred percent certain about the diagnosis when observing the animals. Subject 5 thought the thinking aloud was very useful, because it is easier for the ones around you to understand what you mean. Subject 6 said it was not hard to provide the open answer to the probability questions.

Answering on an answer scale

Subject 1 said he had apparently misinterpreted the scale, because he had only focused on the words on the scale. Subject 2 had, when he first saw the scale, trouble interpreting the scale. The combination of the verbal labels with the numerical labels confused him. Subject 3 thought it was very useful to answer the questions twice, because it makes you think again about the question. Subject 4 indicated that the situations that were presented in the questions were in most cases clear, and therefore it was easy to provide a clear answer. Subject 5 said he had to study the scale for a while before he understood how it worked; he would have preferred an explanation of the scale. Subject 6 said is was not hard to transfer his open answer to the answer scale.

Combination of words and numbers on the scale

Subject 1 thought fifty-fifty is at the correct height. For the other probability words he had an alternative height. 'Almost impossible' relating to the numerical probability of '0' is not possible according to him. You can never say never. It should be approaching '0'.

Subject 2 thought the combination of words and numbers on the scale was confusing. The words 'impossible', 'fifty-fifty' and 'certain' clearly corresponded to '0', '50' and '100'. He added that he thought the verbal label 'expected' was more probable than the verbal label 'probable'. He explained that the layout was confusing, because the labels for 'fifty-fifty', 'impossible' and 'certain' were positioned at the same height as their numerical value, but the other verbal labels are not positioned that precisely. He did not know if that was on purpose, or it was accidentally positioned like this.

Subject 3 would have positioned the verbal labels a bit differently, but thought the labels fit quite well. He thinks the percentages can be better used than the words. The meaning of the words is too imprecise, while numbers have a precise value.

Subject 4 thought the combination of words and numbers made it extra clear what is meant by '100'. It prevents misunderstandings. The word 'uncertain' is unclear he thought, because it means 'it is uncertain there is a relation', but how uncertain is it?

Subject 5 mentioned he did not look very closely at the words. He thought it was confusing that the verbal labels 'certain' and 'uncertain' are in the list. You would think these are opposites, but on the scale they are not. He only looked at the numbers and thinks a scale with only numbers would be more pure. It would be good to add a percentage sign to the numbers.

Subject 6 says he rather thinks in terms of 0 to 100% or 0 to 10. Except for the words 'fifty-fifty', 'impossible' and 'certain', he thinks the words are not at the right height. He had mainly looked at the numbers while using the scale.

Chapter 6: Revisiting the hypotheses

In this chapter the hypotheses are investigated with the results of Experiment 1. The hypotheses are discussed one by one. Some hypotheses are answered with the results, for the investigation of other hypotheses more information is needed. In section 6.1 Hypotheses 1 and 3 are investigated. The results for Hypothesis 3 are discussed in 6.2. Section 6.3 investigates Hypothesis 4, section 6.4 investigates Hypothesis 5, section 6.5 investigates Hypothesis 6 and section 6.6 displays the results for Hypothesis 7. Section 6.7 describes why a second experiment was setup to further investigate the representation of the scales.

6.1. Hypotheses 1 and 2

This section describes the result for Hypotheses 1 and 2. The investigation of these hypotheses is tangled up, therefore they are discussed together in this section. Before we discuss the results, we repeat the hypotheses under consideration.

Hypothesis 1: *The affordance of the original scale does not attract users to indicate their answer as a point-based answer on the vertical line*

Hypothesis 2: *The affordance of the original scale does attract users to indicate their answer on one of the available anchors.*

The hypotheses consider only the affordance of the original scale, and not the affordance of the alternative scale. Nevertheless, the results from all subjects is used to investigate the hypotheses.

The Subjects 2, 4 and 6 were presented with the original version of the scale. In Part 2 of the experiment, none of them indicated their answer as a point on the vertical line (see Section 5.2.1). All three of them indicated their answers by drawing circles around numerical labels (Method 2) or around a combination of a numerical label and a dash on the vertical line (Method 3). All answers of Subjects 2, 4 and 6 are indicated on an anchor. Only one answer is indicated as a rather large circle around the numerical label 100, the dash on the vertical line and a rather large part of the vertical line itself. Because this circle is covering many parts of the scale it is not perfectly clear if this answer is indicated on an anchor.

The Subjects 1, 3 and 5 used the alternative version of the scale. Section 5.2.1. describes that these subjects a total of 9 times indicated their answer as a point on the vertical line. Only 3 answers were indicated on an anchor on the scale. They indicated 4 answers as a range on the scale.

Conclusion

Without explicit instruction the affordance of the original scale does not attract users to indicate their answer as a point on the vertical line of the scale. The fact that all subjects circled a numerical label on the original scale illustrates that the middle column with the vertical line and the horizontal dashes is not perceived as the place to indicate an answer. Drawing a line or a cross is not perceived as the method to indicate an answer. The affordance of the original scale did suggest the users to indicate their answer on one of the available anchors. It seems that the alternative scale does attract subjects to indicate their answers as a point on the scale. Furthermore the alternative scale does not seem to attract subjects to indicate their answer on one of the available anchors.

Discussion

As stated in Section 5.1.1, the subjects who used the original scale had more (10 answers) verbal open answers than the subjects who used the alternative scale (3 answers). Could this have influenced the method and place of indication of their answers on the scale? From the think-aloud of Part 2 and 4 it appears that the perception of the scale caused the subjects to only use the anchors and encircle them. It is possible that the subjects with the verbal open answers would easier agree with a given label, since they do not have an exact number as written open answer to indicate on the scale.

6.2. Hypothesis 3

This section describes the results for Hypothesis 3. Before we discuss the results, we repeat the hypothesis under consideration.

Hypothesis 3: *The alternative representation of the verbal-numerical scale results in less wrongly indicated answers and answers indicated on anchors than the original representation.*

Our experiment provided no instruction about how to indicate an answer on the scale. We wanted to observe the natural manner to indicate answers on the scales. If a scale has a good representation, it should attract users to intuitively use that scale as was intended. As a measure of the affordance of the scale we will count the number of wrongly indicated answers. We defined a wrongly indicated answer earlier as an answer that was indicated in another way than was intended by the designers of the scale. This means that an answer that is not indicated as a point on the vertical line of the scale is a wrongly indicated answer.

<u>Original scale</u>

On the original representation all 17 answers (100%) are wrongly indicated by the subjects. All answers were indicated by circling numerical labels. Sixteen answers were indicated on an anchor. One answer was indicated by circling a numerical label and some other parts of the scale. It is therefore classified as a wrongly indicated answer, but not as an answer indicated on an anchor.

Alternative scale

In the alternative representation one subject indicated all his answers by putting a cross on the vertical line. Another subject indicated four of his answers with an arrow or dash towards one point on the vertical line. This is judged as an answer that is indicated as intended. One subject indicated his answers with curly braces and by circling a numerical anchor. A total of 7 (of a total of 16 answers) answers are classified as wrongly indicated answers. Only 3 of these 7 answers were indicated on an anchor.

Conclusion

The use of the original scale resulted in more wrongly indicated answers (17 answers) than the use of the alternative scale (7 answers). Furthermore, the number of answers that are indicated on an anchor is higher on the original version of the scale (16 answers) than on the alternative one (3 answers).

Discussion

Some information about the process of indicating the answer can be retrieved from the think-aloud of Part 2 and 4. The subjects circled the numerical labels because they thought the existing numerical labels were the only possible answers they could indicate on the scale. Our results provide no explanation for the fact that subjects used the original scale as a 7-point scale, while the alternative scale was mostly used as continuous scale.

6.3. Hypothesis 4

This section describes the results for Hypothesis 4. Before we discuss the results, we repeat the hypothesis under consideration.

Hypothesis 4: *The semantics of the verbal labels 'certain (almost)' and 'uncertain' causes users to indicate how certain they are themselves on the scale instead of indicating the answer to the question.*

This hypothesis was not explicitly tested in the experiment. We hoped the think-aloud would provide us with the answer. Since the verbal labels under consideration are used in both versions of the scale, we will use the results from all subjects. As is stated in Section 5.2.1 five of the six subjects started Part 2 by indicating how certain they were about their answers, instead of indicating the answers itself on the scale. It is not clear from the experiment why they did this. We think it happened because the subjects had to answer the questions in Part 1 already. They might have not understand why they should answer the same question again. From the instruction to answer the question again, in combination with the verbal labels 'certain (almost)' and 'uncertain' on the scale subjects could have concluded they had to fill in how certain they were about the answer given in Part 1 of the experiment. This is an assumption and cannot be derived from the results from the experiment. During the experiment the subjects did have some comments on these two specific verbal labels (see Section 5.3.1). The words 'certain' and 'uncertain' do cause some confusion. The words seem to be opposites in their meaning but they are not opposites in the scale. Furthermore

these two labels are meant to say something about the probability of the posed question, instead some subject use the words 'certain' and 'uncertain' to talk about themselves.

Conclusion

The verbal labels 'certain' and 'uncertain' caused for two subjects some confusion about the interpretation of the scale. We do not believe that these two labels caused subjects to indicate how certain they are instead of indicate their answer.

6.4. Hypothesis 5

This section describes the results for Hypothesis 5. Before we discuss the results, we repeat the hypothesis under consideration.

Hypothesis 5: *When asked for a probability assessment, users provide an answer in a format different from a point estimate.*

In Part 1, five subjects provided their written open answer in a format different than a point estimate. One subject (Subject 3) answered with numerical point estimates. Most subjects expressed their answer as a numerical range or as a verbal probability.

The think-aloud provides some clues about the reason for why Subject 3 did answer the probability questions with a numerical point estimate. In relation to the other subjects, Subject 3 provided his answer rather quickly, without much thinking aloud. He also seemed very confident about his answers. He did not question the meaning of the terms or the amount of information that was provided in the questions and he made no reservations in his answers.

Conclusion

When asked for a probability assessment, 5 out of 6 users provide an answer in a format different from a point estimate.

Discussion

It is possible that the instruction to “answer the question like you would to a fellow veterinarian” has encouraged some subjects to provide a verbal open answer. From the results there is no sign that the instruction did influence the type of answer. This instruction was given, because of an observation during a practice run of this experiment. One subject provided numerical point estimates for the sole reason that he thought the elicitor had to calculate with it later on. The subject's answer was influenced because he adapted his answer to the attending elicitor.

6.5. Hypothesis 6

This section describes the results for Hypothesis 6. Before we discuss the results, we repeat the hypothesis under consideration.

Hypothesis 6: *The combination of probability words with probability numbers in the verbal-numerical scale causes confusion about the use of the scale.*

Since both the original and the alternative scale combine probability words with probability numbers, the results from all subjects are taken into account. The subjects did have some remarks about the combination of the probability words with the probability numbers. The remarks mostly covered the combination of specific words with specific numbers. The subjects agree on the position and value of the words 'impossible', 'fifty fifty' and 'certain'. Subject 2 said to be confused by the combination of the words and numbers. The number he wanted to indicate on the scale leads him to a corresponding verbal label that does not want to indicate on the scale. Furthermore he thinks the order of the verbal labels is not correct. This positioning of both the verbal and the numerical labels on the original scale he used contributed to the confusion. Subject 4 however thinks the combination makes it extra clear what is meant with the numbers. All subjects indicated they would rather use numbers than words in the scale, because numbers are more precise.

Conclusion

The method of combining probability words with probability numbers does cause confusion about the use of the scale for one subject. He felt that the verbal label that was positioned next to the number he wanted to indicate was not the right verbal value. He therefore seemed confused if he should indicate the numerical label or the verbal label he thought was the correct answer. The subjects do comment on the use of specific verbal labels and comment on the positioning of the verbal and numerical labels. Although they have some remarks, they do not have observable problems with the use of the scale.

6.6. Hypothesis 7

This section describes the results for Hypothesis 6. Before we discuss the results, we repeat the hypothesis under consideration.

Hypothesis 7: *Users adjust their answer to fit an existing anchor.*

We investigated this hypothesis by comparing the written open answers from Part 1 with the answers indicated on the scale in Part 2. If the answer from Part 1 differs from the answer in Part 2, the think-aloud from Part 2 and 4 explains the reason for this difference. This difference is important because both answers are answers to the same probability question. Several reasons were found to explain the differing answers. We determined that some subjects did adjust their answer to fit an

existing anchor on the scale. In Section 5.2.2. we investigated per subject the reason for differences between the written open answer and the answer on the scale.

The subjects who used the original scale all adjusted their answers towards an existing numerical anchor. This is true for the answers that did not have already the same value as the indicated numerical label. Two subjects said they rather would have indicated another number, but thought the available numerical anchors were the only possible answers. The remarkable thing is that they do accept the interpreted answer on the scale from the elicitor.

From the subjects who used the alternative scale 2 subjects adjusted their answers. One because he did had only looked at the verbal labels on the scale, and one because he had come upon new insights by examining the question for the second time. So none of the subjects who used the alternative scale adjusted their open answers towards an existing anchor.

Conclusion

The subjects using the original scale adjusted their answers towards existing anchors. But they consider both the written open answers as the answers on the scale as a correct answer. The subjects using the alternative scale did not adjust their answers towards anchors on the scale.

6.7. A second experiment

From the results from Experiment 1 we conclude that the original and the alternative representation of the verbal-numerical scale both attract different use of it. Several changes were made at once to the original scale, therefore the individual influence of the changes cannot be traced. A second experiment was setup as a means to investigate what changes to the original scale is causing the subjects to use it differently.

Chapter 7: Method Experiment 2

In Experiment 1 we saw that the subjects using the original scale acted differently than the subjects that used the alternative scale. To investigate the cause of these differences we set up a pilot experiment. We wanted to look at the scale by going back to the basics of perceiving and interpreting objects like scales. This pilot experiment presented subjects with several images of both the original and the alternative representation of the verbal-numerical scale. Subjects were asked to group parts of the scale which they think belong together by encircling them.

In the first section the literature on the basics of object perception is outlined. Section 7.2. describes the process of perception of the verbal-numerical scale. Section 7.3. contains the problem description. The design of the pilot experiment is described in Section 7.4. Section 7.5 outlines the experimental set-up and materials. The participants are discussed in the last section.

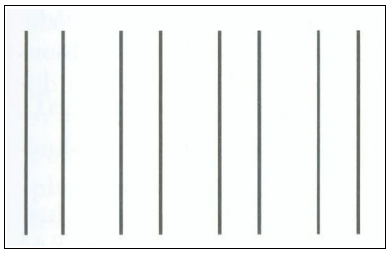
7.1. Object perception

When people notice something, they first perceive it as segmented objects like lines, bars and dots [19]. This first perception is however not sufficient to interpret what is seen. Therefore people need to know which segmented objects belong together and form a unit. This grouping of objects provides support in giving meaning to what is seen; to interpret the world. People tend to organize objects according to the gestalt principles of organization. Furthermore, experience plays a role in grouping objects. In 7.1.1. the gestalt principles of organization are described and illustrated. In 7.1.2. the role of our experience is explained in grouping objects.

7.1.1. Gestalt principles of organization

According to Wertheimer [20] the form and organization of objects support people to organize objects into units if they are close together (proximity), similar to one another (similarity), form a closed contour (closure), move in the same direction (good continuation), are located within the same perceived area (common region) or are perceived in any uniform, connected region (connectedness) [21]. These characteristics are called the gestalt principles of organization. The idea

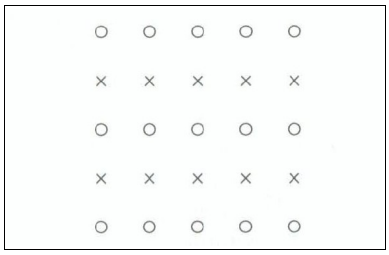
of the underlying theory of the gestalt principles is that the whole of the objects is more than the sum of its parts [22]. Below we describe the principles and illustrate them.



Proximity

Figure 7.1. consists of eight vertical lines, but they are not perceived as eight separate lines. People tend to perceive them as four pairs of lines instead. The lines that are close to each other are grouped into units.

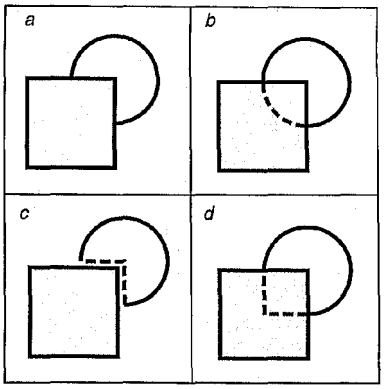
Figure 7.1. Illustration of gestalt principle 'proximity' (source [19])



Similarity

Figure 7.2 consists of O's and X's. Rows of O's and rows of X's are perceived. This illustrates the principle of similarity. Objects that look alike are grouped into units, in this case in rows.

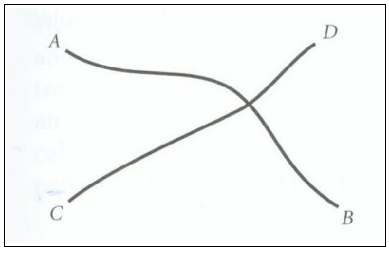
Figure 7.2. Illustration of gestalt principle 'similarity' (source [19])



Closure

In Figure 7.3 people see a circle that is occluded by a square (b). The occluded object can however have many other possible shapes than a circle (c and d).

Figure 7.3. Illustration of gestalt principle 'closure' (source [19])



Good Continuation

In Figure 7.4 two lines are seen: one from A to B and one from C to D. There is no reason why the Figure would not display a line from A to C and another from B to D. The perceived lines however show better continuation, while the alternative lines have a sharp turn in the line.

Figure 7.4. Illustration of gestalt principle 'good continuation' (source [19])

Common Region

Figure 7.5 consists of eight black dots and four rectangles. The rectangles each seem to contain two black dots. The two dots in each rectangle are perceived as objects in the same region and are grouped into one unit.

Figure 7.5. Illustration of gestalt principle 'common region' (source [21])

Connectedness

Figure 7.6 consists of eight black dots and four lines. These objects are perceived as four units each consisting of two connected (by a line) black dots. There is a strong tendency to perceive any uniform, connected region as a single unit.

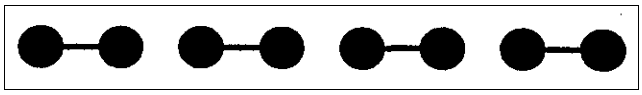
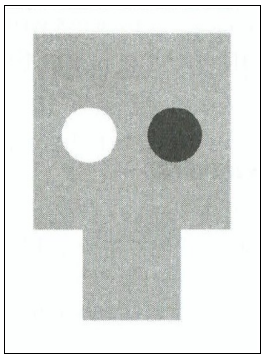


Figure 7.6. Illustration of gestalt principle 'connectedness' (source [21])

7.1.2. Experience

According to Arnheim [15], what one thinks an image represents depends on what it reminds one of. Past experiences come up at the moment something is perceived. Experience is different for every individual. A person's experience depends for instance on his age, his social background, the place he grew up and his education. To illustrate how experience plays a role in perception, consider Figure 7.7 below.



What is it? Is it the head of a duck with one eye open and one eye closed? Is it a block with holes in it, an usb-stick with two dots? What you think it is, depends on the past experiences you bring to the moment you perceive the image.

Figure 7.7. Illustration of the role of experience in object perception (source [16])

7.2. Perception of the verbal-numerical scale

When looking at the verbal-numerical scale the segmented objects are the first things that are noticed by the user. Users will see that the scale consists of strings of tokens (for example “waarschijnlijk” and “50”) and lines (horizontal and vertical). The segmented objects will unconsciously be grouped into units according to the gestalt principles of organization. For instance, the horizontal lines could be grouped with the vertical line on the basis of the principle 'connectedness' (Figure 7.8.a). On the basis of the gestalt principle 'proximity', the strings of tokens

on the right side could be grouped with the horizontal lines that are positioned close to them (Figure 7.8.b). The gestalt principle 'similarity' could lie on the basis of grouping all labels on the right side together, because of their similar size (Figure 7.8.c).

Experience will also play a role in grouping the objects of the verbal-numerical scale. For a scale to be perceived as a scale it has to remind the user of scales that he has seen before. In a later stage the strings of tokens are recognized as words and numbers. Users will then group objects on the basis of their semantics. For instance, the word 'fifty-fifty' could be grouped with the number '50', because the semantics are similar (Figure 7.8.d).

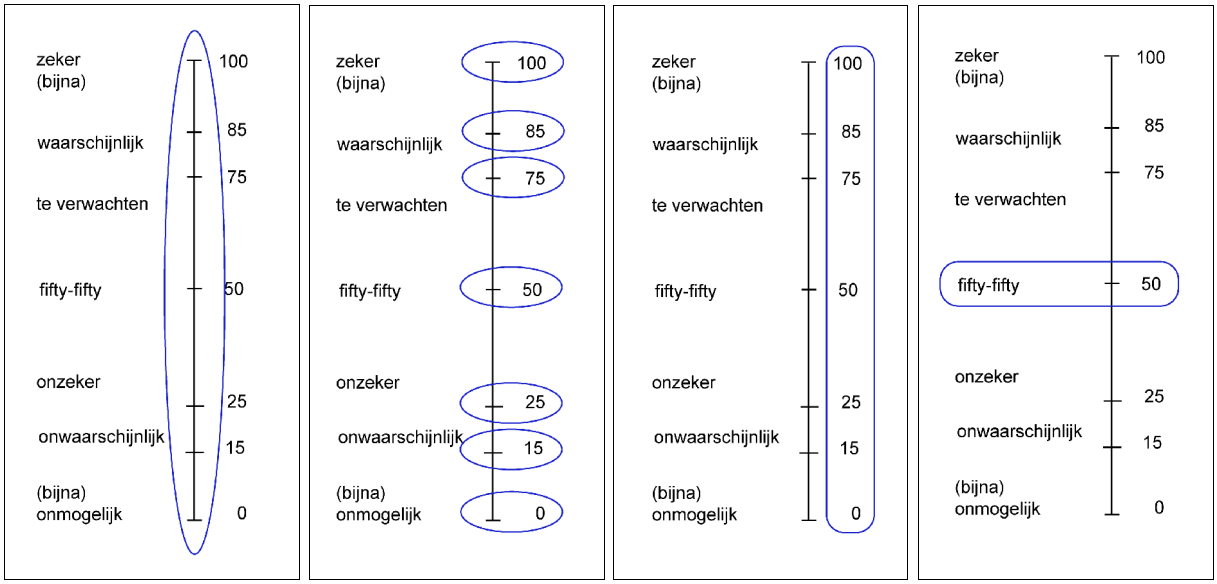


Figure 7.8.a **Figure 7.8.b** **Figure 7.8.c** **Figure 7.8.d**

Figure 7.8. Illustration of grouping on the basis of gestalt principles 'connectedness' (a), 'proximity' (b), 'similarity' (c) and on the basis of semantics (d).

7.3. Problem description

In the first experiment, subjects using the original scale only encircled numerical labels to indicate their answer. This resulted in a limited set of indicated answers from the subjects. Subject using the alternative scale indicated their answer mainly as a point on the vertical line by drawing a cross or an arrow. The variety of the answers from those subjects was therefore much greater. It seems that the original scale was interpreted as a limited set of answers that has to be chosen from, while the alternative scale was interpreted as a continuous scale on which the subject is not bounded by a limited set of answers.

We want to know what the cause is of the different behavior of the subjects on the two scales. We assume that the changes that were made during the design of the alternative scale are responsible for the difference in behavior. Several changes were made at once, such as adding numerical labels and aligning the verbal labels towards the vertical line, therefore the influence of the individual changes could not be traced.

7.4. Design

In the second experiment we investigate the subjects grouping behavior on 18 different scale images. The scale images vary in three aspects: their organization, the extent of looking like a scale and the level of their semantic content. By asking the subjects to group objects in the scale images, we aim to capture on paper the grouping process that takes place in the subjects mind when presented with an image. Section 7.4.1. lists the variables of the pilot experiment. The design of the scale images is outlined and illustrated in Section 7.4.2.

7.4.1. Variables

The independent variables of the pilot experiment are:

O - Organization: the position of the objects in relation to each other. Two conditions:

1. organization of the original verbal-numerical scale
2. organization of the alternative verbal-numerical scale

R – Resemble scale appearance: to what extent the image looks like a scale. Three conditions:

1. both vertical line and horizontal lines
2. only horizontal lines
3. no lines at all

S - Semantics: to what extent semantics are present. Three conditions:

1. semantics: probability words and numbers
2. semantics: unrelated non-probability words and numbers
3. no semantics: only black rectangles

The dependent variables are:

1. the clusters that are made by the subjects
2. the motivation for the grouping

7.4.2. Design of the scale images

A total of 18 scale images were designed. In this section the design of the scales for the different conditions of the variables are described and illustrated.

Variable 1: Organization

The organization of the objects, their relative position to each other, is the basis for grouping the objects into units. We want to investigate if subjects make different units in the scale images with the organization of the original scale than in the scale images with the organization of the alternative scale. As a reminder the original scale and the alternative scale are depicted in Figure 7.9.

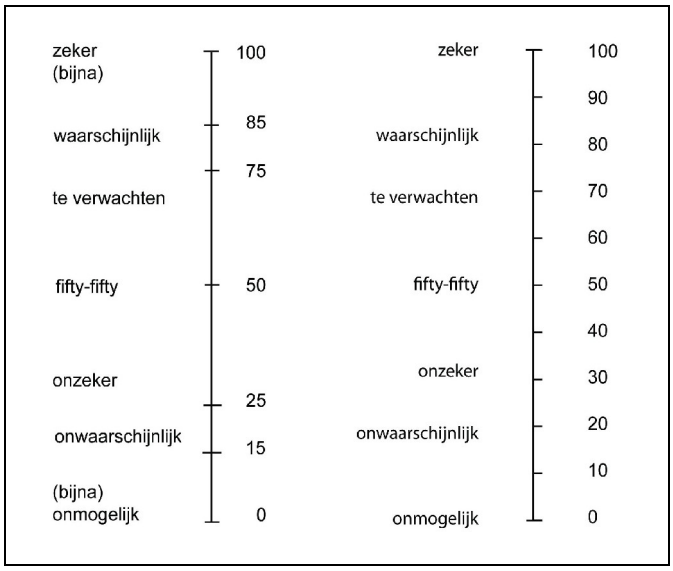


Figure 7.9. The organization of the original verbal-numerical scale (left) and that of the alternative verbal-numerical scale (right).

Variable 2: Resemble scale appearance

We hypothesize the horizontal lines in combination with the vertical line are recognized as a scale because people have seen such scales before. Their experience with other scales leads them to group the vertical line with the horizontal lines. By designing scale images with varying presence of the vertical line and the horizontal lines, we test whether the presence of these objects influences the kind of units the subjects make.

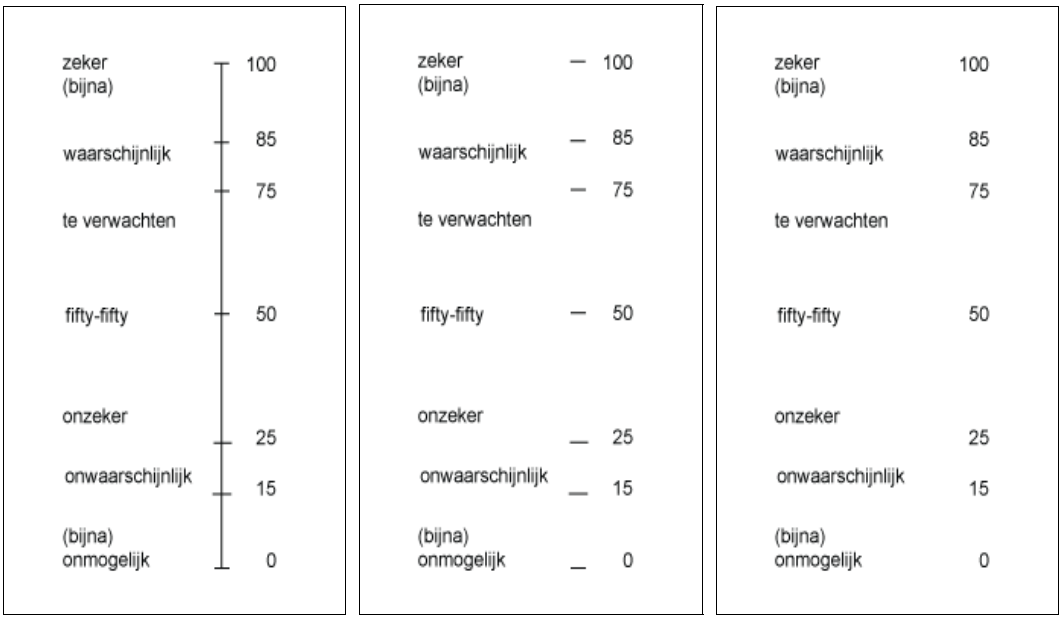


Figure 7.10. Scale images in the three conditions of the variable 'scale likeness (experience)': both vertical line and horizontal lines (a), only horizontal lines (b) and no lines at all (c).

Variable 3: Semantics

We know that the semantics of the objects are processed in a later stage. In order to investigate the grouping of objects on the basis of their form and organization, the subjects should not be distracted by the semantics of the objects. Therefore we designed scale images in which all (semantic) labels

are replaced with black rectangles of the same size as the original label (Figure 7.11.c). To investigate the influence of the semantics of the verbal-numerical scale we also designed scale images in which the verbal labels are replaced with unrelated non-probability words, like 'flower' and 'turtle' (Figure 7.11.b).

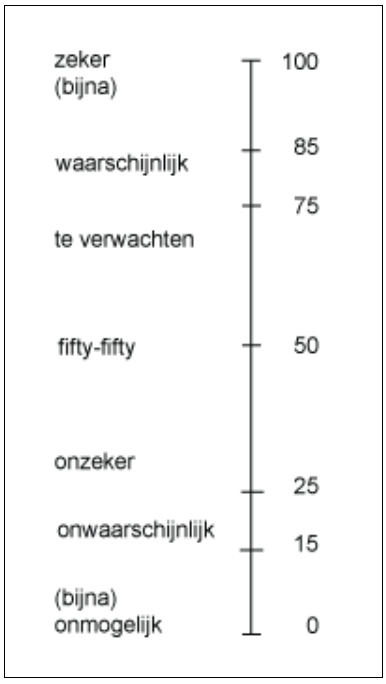


Figure 7.11.a.

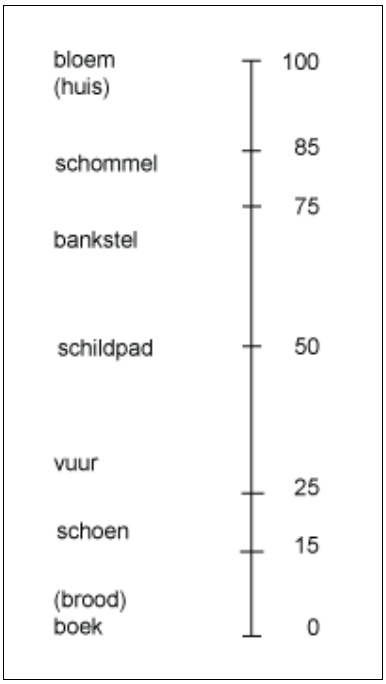


Figure 7.11.b.

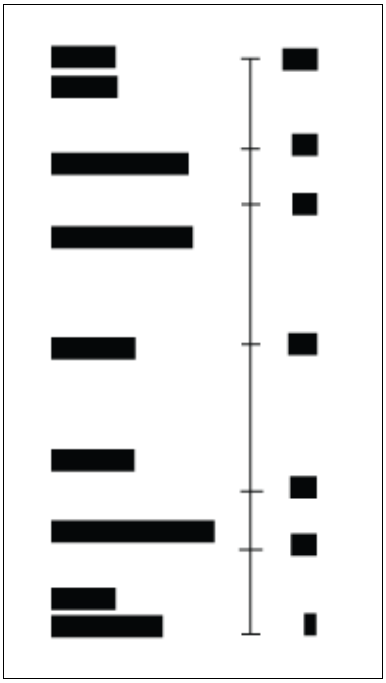


Figure 7.11.c.

Figure 7.11. Scale images in the three conditions of the variable 'semantics': probability words with numbers (a), unrelated non-probability words with numbers (b), black rectangles of the same size as the original label (c).

To distinguish the scale images and the sheets we tagged every sheet with a code. An example is shown in Figure 7.12. The first part of this code refers to the Subject and sheet order. The second part (after the "/") refers to the variable conditions of the scale image. The coded sheets can be found in Appendix G.

Sheet 4.2/O1-S3-R3 (Figure 7.12) is:

- the second sheet that Subject 4 examined (4.2)
- the first condition of the variable Organization, being the original organization (O1)
- the third condition of the Variable Semantics, being the black rectangles (S3)
- the third condition of the Variable Resemble scale appearance, being a scale image without any lines (R3)

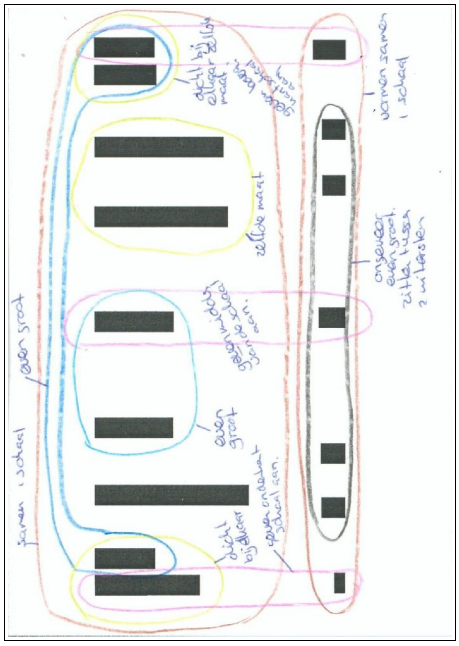


Figure 7.12. Sheet 4.2/O1-S3-R3

7.5. Experimental setup and materials

The experiment is divided in three parts. The whole experiment is in Dutch.

- Part 1: short introduction with an example of clustered objects (see Figure 7.13)
- Part 2: subjects group the objects they think belong together in three scale images
- Part 3: subjects motivate why they made every group

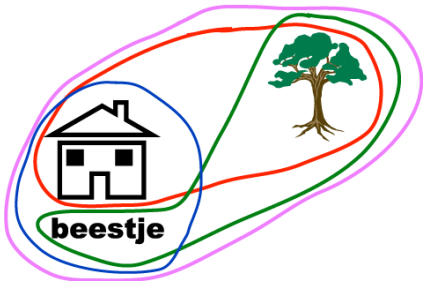


Figure 7.13 sheet of Pilot Experiment with an example of grouped objects.

For every subject a set of scale images was prepared at A6-size. Every set contained the assignment (Appendix H), an example image (figure 7.13), and three assignment pages with a scale image. The assignment was to cluster the objects which they think belong together by encircling them. Between the assignment empty sheets were inserted, to prevent the subject from looking at the next assignment page. Between each set of scale images the conditions of the variable 'organization' are alternated. The first two sheets contained always a scale image in the non-semantic condition (S3). The third sheet always contained a scale image in a semantic condition (S1 and S2). The intention was to ensure that the subject would not be influenced by the semantics of an earlier sheet while assessing a non-semantic scale image. Furthermore the subjects were provided with colored pencils for drawing circles on the assignment pages.

The experiment is conducted in a one-to-one setting with the researcher and the subject in a quiet area. To start, the subject is told to read the assignment on the first page. Furthermore the term 'object' is specified. If necessary the assignment and the example are further explained. The subject is instructed to start the experiment and told that it is not allowed to revisit pages that are already flipped. After all scale images are assessed by the subject, the researchers asks for a motivation for every grouping that has been made. The researcher wrote this motivation down in keywords. After the experiments the researcher explains to the subjects what has been measured in the experiment and why this experiment is conducted. All subjects were surprised with a drink afterward.

7.6. Participants

Nine Dutch speaking subjects participated in this experiment. The demographic characteristics of these subjects are not considered relevant. The age of the participants ranged from 20 – 40 and there were both male and female subjects who participated. The educational background from the subjects varied from vocational education and university.

Chapter 8: Results Experiment 2

The results of Experiment 2 are presented in this chapter. A total of 9 subjects assessed a total of 27 sheets with one of the 18 designed scale images on it. Because of the small group of subjects and the order of the sheets several scale images have been assessed more times than others. We will not measure the results between and within subjects, but treat the sheets as one group.

The organization of this chapter is a follows. First, it is outlined on what measures the variables will be compared. In Section 8.2. the results are presented and possible explanations for these results are given. Section 8.3 lists additional remarks about the experiment and Section 8.4. contains the conclusion of this Pilot Experiment.

8.1. Cluster and motivation types

By comparing the sheets of the subjects we looked for interesting patterns. We counted the occurrence of three types of clusters and three types of motivations for the clusters; these cluster types and motivation types are described in this section. All sheets that are mentioned as example can be found in Appendix I.

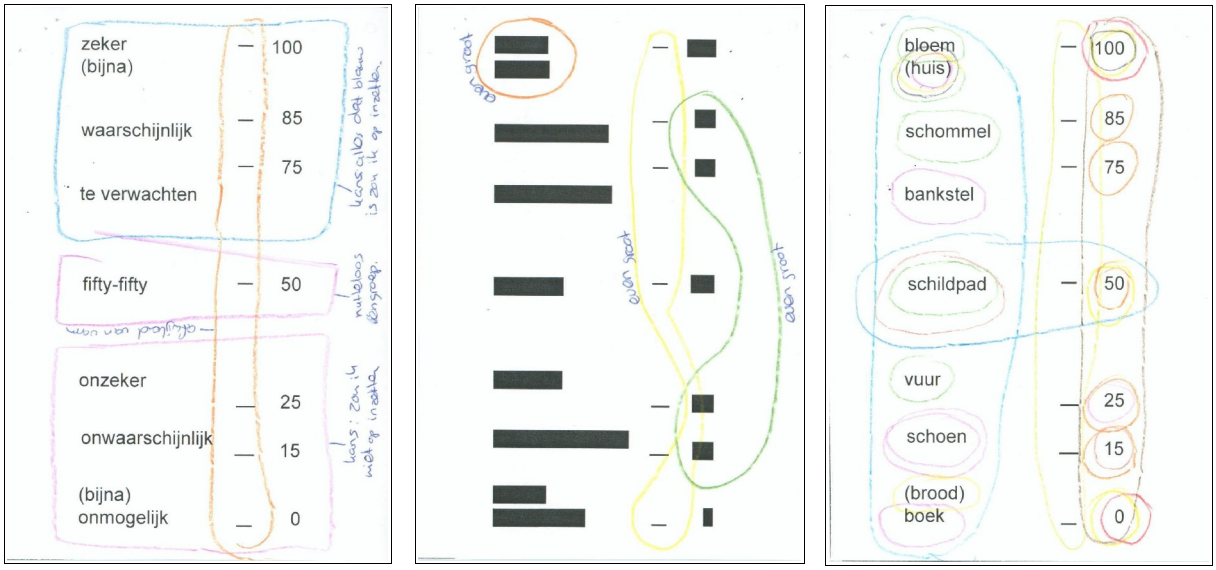
Cluster types

From analyzing the sheets of the subjects we distinguish three types of clusters. These types are listed below and illustrated with an example figure.

1. Horizontal clusters: clusters that include objects from more than one column, for example the sheet in Figure 8.1a has 3 horizontal clusters.
2. Vertical clusters: clusters that capture at least three united objects from one column, for example the sheet in Figure 8.1b has 1 vertical cluster (in the middle column).
    - Column cluster: vertical cluster that includes one or more complete columns, for example the sheet in Figure 8.1c has 3 column clusters.

The categorization is not complete, because not all clusters can be categorized in one of the above defined categories. For instance the green cluster in Figure 8.1b that includes two pairs of black rectangles from the right column is not categorized as a vertical cluster because it does not contain three united objects from one column. We believe that a cluster with three united objects form one column shows an intention of a vertical cluster. Some clusters belong to more than one category, for

instance the blue cluster at the top of the sheet in Figure 8.1a. This cluster is both a horizontal cluster and a vertical cluster.



a. Sheet 3.3/O1-S1-R2 **b. Sheet 1.2/O1-S3-R2** **c. Sheet 9.3/O1-S2-R2**

Image 8.1. Sheets as illustration of the three categories of clusters.

Motivation types

The motivation is the reason that is provided by the subject of why a certain cluster of objects was made. We distinguish three categories of motivations:

1. Semantic motivation: from this motivation it be clear that the subject had semantic reason for the cluster. Semantic motivations are for instance “schildpad kan 100 jaar worden” (In English: a turtle can turn a hundred years old) as in sheet 8.3/O1-S2-R1 or “kans zou ik niet op inzetten'' (In English: these are chances I would not bet on) as in sheet 3.3/O1-S1-R2.
2. Positional motivation: from this motivation it is clear that the position of the objects, the layout, is the reason to cluster. Positional motivations are for instance “staan rechts” (In English: are positioned on the right) as in sheet 3.2/O2-S3-R2 or “bovenste” (In English: at the top) as in sheet 9.2/O2-S3-R3.
3. Form motivation: from this motivation it is clear that the form or shape of the objects is the reason to cluster them. Form motivations are for instance “afwijkend van vorm” (In English: dissimilar form) as in sheet 3.2/O2-S3-R2 or “allemaal even groot” (In English: all the same size) as in sheet 1.1/O2-S3-R3.

Three motivations cannot be categorized in one of the above defined categories. These three will be listed below.

- “alsof ze elkaar aan zouden vullen” (In English: like they would complete each other) in the sheet in Figure 8.2a. It is not clear why (form, position) the clustered object would complete each other. Therefore this motivation cannot be categorized.
- “is eigenlijk 1 item” (In English: is actually one item) in the sheet in Figure 8.2b. From this motivation it is also not clear why (form, position) the subject considers the subjects as one thing.

- “hij vertegenwoordigt in zijn eentje die” in the sheet in Figure 8.2a. From the sheet it is clear that he clusters one object from the left column with three objects from the right column because the first one represents the other three. It is not known on what basis the subject assumes this (form, position).

a. Sheet 9.2/O2-S3-R3 **b. sheet 4.1/O2-S3-R1**

Figure 8.2. Sheets as illustration of motivations that cannot be categorized.

In few occasions a cluster can have more than one motivation, for instance “2 grote, 2 kleine, zelfde lijn” (In English: 2 large ones, 2 small ones, on the same line) in 9.2/O2-S3-R3. The first part of the motivation is about the size of the objects and therefore categorized as a form motivation. The second part is about the fact that the objects are positioned on the same line and therefore categorized as a position motivation.

8.2. Results

In this section the variables are compared and the results are explained. In Section 8.2.1. the conditions of the Variable Organization are compared. Next, the results of the Variable Semantics are presented in Section 8.2.2. The conditions of the Variable Resemble scale appearance are compared in Section 8.2.3. In Section 8.2.4. individual interesting clusters are described.

8.2.1. Organization (O)

We compared the two conditions of the Variable Organization (O): the Original Organization (O1) and the Alternative Organization (O2). Table 8.1. shows the total number of sheets of each condition and the total number of clusters that were made. To compare the total number of clusters, the average was calculated. Table 8.1. shows that more clusters were made in O1 than in O2.

	Organization	
	O1	**O2**
Total sheets	14	13
Total clusters	90	76
Cluster average	**6,4**	**5,8**

Table 8.1. The number of sheets, the number of clusters and the averged number of clusters in the two conditions (O1-original and O2-alternative) of the Variable Organization (O).

In Table 8.2. the occurrence of the cluster and motivation types is compared in the Original Organization (O1) and Alternative Organization (O2). To compare the amount of clusters occurring in each type, the numbers are expressed as percentages of the total number of clusters in each condition.

	Organization			
	O1		**O2**	
Total number of clusters	90	100,00%	76	100,00%
CLUSTER TYPES				
Horizontal clusters	19	21,00%	56	73,00%
Vertical clusters	20	22,00%	21	27,00%
Column clusters	12	13,00%	14	18,00%
MOTIVATION TYPES				
Semantic motivation	23	25,00%	20	26,00%
Positional motivation	15	16,00%	45	59,00%
Form motivation	46	51,00%	20	26,00%

Table. 8.2. The percentage of the total number of clusters and motivations in the two conditions (O1-original and O2-alternative) of the Variable Organization (O).

Below the results are listed that can be derived from Table 8.2 followed by a possible explanation of that result.

Result 1: In Condition O2 (alternative representation) a dramatically higher percentage of horizontal clusters were made than in Condition O1 (original representation).
Explanation: In Condition O2 the columns are positioned closer to each other because the columns are centered towards the middle. Therefore the objects are also positioned closer to each other. This is supported by the next result, Result 2.

Result 2: In Condition O2 a positional motivation occurs relatively more often than in Condition O1.
Explanation: In order to cluster objects on a positional motivation, we believe the subject should be aware of the differences and similarities in the position of the objects. We think O2 has a more clear organization with the right and left columns aligned towards the middle. The organization of O1 might look more arbitrary. Subjects could have been more aware of the position of the objects in O2 and therefore had more positional motivations.

Result 3: In Condition O1 a form motivation occurs relatively more often than in Condition O2.
Explanation: It could be that the subjects were more aware of the form of the objects in O1 than they were in O2. We think this is connected with the assumption that the subjects are more aware of the position of the objects in O2 than in O1. The subjects have in O2 more choice to motivate their clusters. Clustering with a form motivation mostly occurs in the condition with the black rectangles (S3). In condition O1 and O2 the same amount of sheets with black rectangles were assessed.

On the other cluster and motivation types no clear differences can be observed.

Conclusion
Results 1-3 leads us to consider that subjects might integrate the different columns better in the alternative organization(O2) than in the original organization (O1). This is because maybe they are more aware of the fact that the objects are organized and not positioned arbitrarily.

8.2.2. Semantics (S)

We compared the three conditions of the Variable Semantics: the probability words with the numbers (S1) the non-probability words with the numbers (S2) and the black rectangles (S3). Table 8.3. shows the total number of sheets of each condition and the total number of clusters that was made. To compare the total number of clusters, the average amount of cluster per sheet was calculated. It is shown that subjects made on average some more clusters in S2 than in S1 and S3. S3 might have less clusters, because semantics are not on discussion in that condition.

	Semantics		
	S1	**S2**	**S3**
Total sheets	5	4	18
Total clusters	31	26	109
Cluster average	**6,2**	**6,5**	**6,1**

Table 8.3. The number of sheets, the number of clusters and the average amount of clusters in the three conditions (S1-S3) of the Variable Semantics (O).

In Table 8.4. the occurrence of the cluster and motivation types compared in S1, S2 and S3. To compare the number of clusters occurring in each type, the numbers are expressed as percentages of the total number of clusters in each condition.

	Semantics					
	S1		S2		S3	
Total number of clusters	31	100,00%	26	100,00%	109	100,00%
CLUSTER TYPES						
Horizontal clusters	24	77,42%	4	15,38%	47	43,12%
Vertical clusters	5	16,13%	7	26,92%	30	27,52%
Column clusters	4	12,90%	4	15,38%	20	18,35%
MOTIVATION TYPES						
Semantic motivation	27	87,10%	17	65,38%	0	0,00%
Positional motivation	3	9,68%	1	3,85%	56	55,96%
Form motivation	3	9,68%	8	30,77%	61	65,59%

Table 8.4. The percentage of the total number of clusters in the three conditions (S1-S3) of the Variable Semantics (S).

Below the results are listed that can be derived from Table 8.4 followed by a possible explanation of that result.

Result 4: S1 shows the highest percentage of horizontal clusters, S3 shows the second most horizontal clusters. In S2 the least horizontal clusters were made.
Explanation: In S1 the left column objects have an intended semantic relation with the right column objects. This is a strong motivation to cluster objects from the different columns. In S3 the left column objects have the same appearance, the same form, as the right column objects, namely black rectangles. This is also a motivation to cluster objects from the different columns. In S2 both form and semantics are less obvious motivations for making horizontal clusters. The left column objects in S2 have no obvious semantic relation with the right column objects. Furthermore the objects from the two columns have a different form. The objects from both the right and the left column are strings of tokens but the semantics divide them in words and numbers, and therefore different forms.

Result 5: In S1 relatively more semantic motivations occur than in S2.
Explanation: See the explanation of Result 4.

Result 6: In S3 relatively more clusters were made with a positional and a form motivation than in S1 and S2. **Explanation**: The conditions S1 and S2 both contain semantic objects. Condition S3 contains only non-semantic objects. The conditions S1 and S2 offer more categories of motivation to cluster objects than S3.

On the other cluster and motivation types no clear differences can be observed.

Conclusion
The Results 4-6 show that semantic motivation seems to overrule positional and form motivation in S2 and even more in S1. It looks like the semantics in S1 and S2 are distracting the subject from looking at position and form. Semantics might be a stronger motivation for clustering objects than form and postion.

8.2.3. Resemble scale appearance (R)

We compared the three conditions of the Variable Resemble scale appearance: the vertical line with the horizontal lines (R1), only the horizontal lines (R2) and no lines (R3). Table 8.5. shows the total number of sheets of each condition and the total number of clusters. To compare the total number of clusters, average amount of clusters per sheet was calculated. It can be seen that subjects made more clusters in R1 and R2 than in R3. This result seems to be obvious, since R1 and R2 consist of more objects than R3. There are just more objects to cluster.

	Variable Resemble scale appearance		
	R1	**R2**	**R3**
Total sheets	8	9	10
Total clusters	59	59	48
Cluster average	**7,4**	**6,5**	**4,8**

Table 8.5. The number of sheets, the number of clusters and the average number of clusters in the three conditions (R1-R3) of the Variable Resemble scale appearance (R).

In Table 8.6. the occurrence of the cluster and motivation types is compared in R1, R2 and R3. To compare the number of clusters occurring in each type, the numbers are expressed as percentages of the total number of clusters in each condition.

	Resemble scale appearance					
	R1		**R2**		**R3**	
Total clusters	59	100,00%	59	100,00%	48	100,00%
CLUSTER TYPES						
Horizontal clusters	27	45,76%	28	47,46%	20	41,67%
Vertical clusters	13	22,03%	16	27,12%	13	27,08%
Column clusters	9	15,25%	12	20,34%	7	14,58%
MOTIVATION TYPES						
Semantic motivation	16	27,12%	15	25,42%	12	25,00%
Positional motivation	20	33,90%	23	38,98%	17	35,42%
Form motivation	18	30,51%	31	52,54%	21	43,75%

Table 8.6. The percentage of the total number of clusters in the three conditions (R1-R3) of the Variable Resemble scale appearance (R).

Below the results are listed that can be derived from Table 8.6 followed by a possible explanation of that result.

Result 7: The percentages of horizontal, vertical and column clusters for R1, R2 and R3 are close to each other.
Explanation: The presence and absence of (some of) the lines did not influence the type of clustering. The organization, semantics and form of the objects are enough reason to make clusters on in all three conditions.

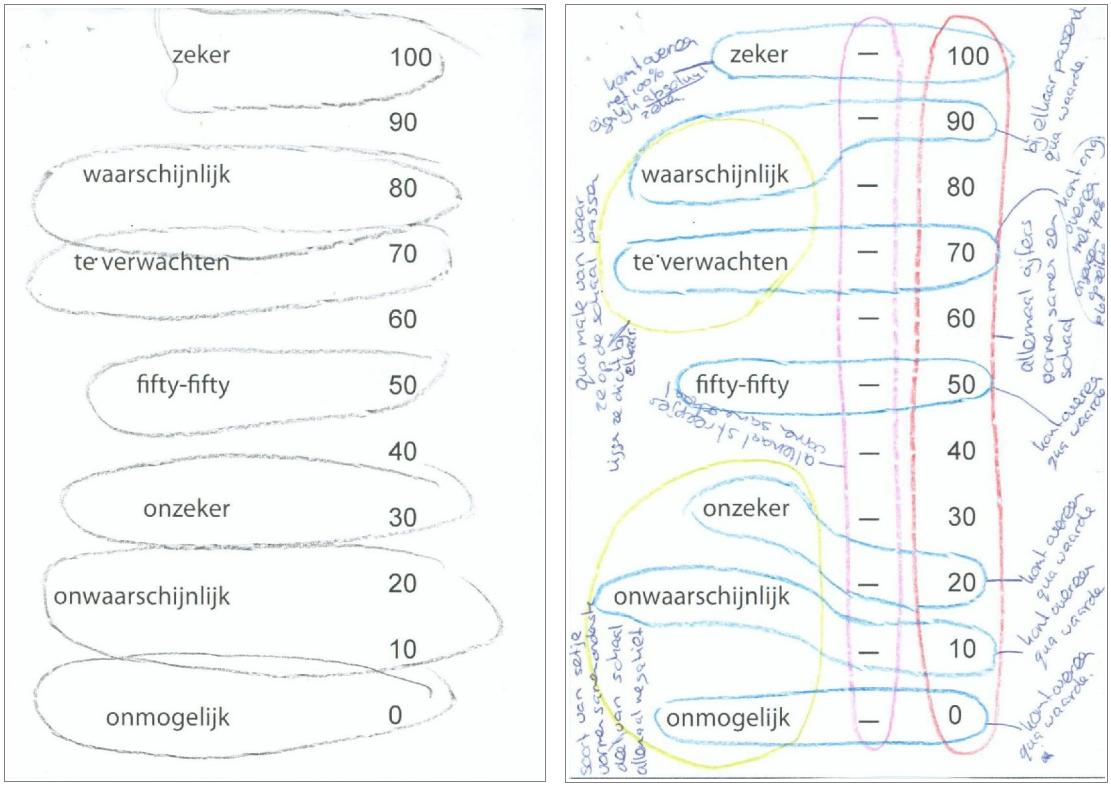
Result 8: In R2 and R3 a higher percentage of form motivations occur than in R1.
Explanation: It could be that the vertical line with the horizontal dashes on it in R1 reminds subjects of a scale. This could have prevented subjects from clustering on the basis of form.

Conclusion
The overall conclusion of the comparison of the conditions of the Variable Resemble scale appearance is that the influence of the presence and absence of (some of) the lines is not big. The differences on the measure points of the three conditions are not big.

8.2.4. Interesting individual clusters

Apart form the comparison of the percentage of cluster and motivation types we observed some interesting clusters and motivation that

In Figure 8.3a the subject has clustered probability words with numbers with the motivation that their values match. This behavior of matching the individual probability numbers with the individual probability words that are positioned at almost the same height was also seen in Experiment 1. Another subject did the same in Figure 8.3b, but deviated one time by clustering the probability word 'onzeker' (In English: uncertain) with the number '20' which are not positioned on the same height. The subject thought the semantics were more important than the positioning.



a. Sheet 5.3/O2-S1-R3 **b. Sheet 4.3/O2-S1-R2**

Figure 8.3. Two sheet illustrating interesting individual clusters and motivations

Subject 1 has clustered 'fifty-fifty', 'onzeker' (in English: uncertain) and '50'. The motivation was “fifty-fifty dan ben je onzeker 50%” (In English: fifty-fifty than you are uncertain 50%). This is

interesting, because this subject gives 'fifty-fifty', 'uncertain' and '50' the same meaning, the same value. Moreover he connects the word 'uncertain' to the person who is answering the scale, by saying 'you are uncertain' instead of 'it is uncertain'. The cluster with this motivation is found in Appendix I (1.3/O2-S1-R1).

8.3. Remarks

When afterwords the goal of the experiment was explained, only one participant mentioned she had recognized a scale in the sheets. We suspect this subject (4) has overheard another subject talk about a scale, because her behavior was very different than that of the other subjects. She talked about a scale immediately when she started with the first sheet. It is possible that she is the only subject who recognized a scale on her first sheet (which showed both the vertical line and the horizontal dashes) and realized that the other two sheets also represented a scale. Her sheets can be found in Appendix I. Another participant had only recognized a scale in the meaningful version of the scale, on the third sheet. The other participants were surprised that the experiment was about a scale. They did not realize all the objects belonged together on one sheet. Also they had not recognized the similarity between the three different sheets. They thought every sheet was a complete new image to them. Even the subjects who were presented with a first and third sheet that were completely similar in their organization (like sheet 1 and 3 in Image 8.4), but differed in the semantics of the objects (black rectangles in the first sheet and words and numbers in the third) had not recognized the two sheets were related.

It was also interesting that all of the participants were anxious about the meaning of the clusters they made. They thought they would lead to a psychological description of themselves.

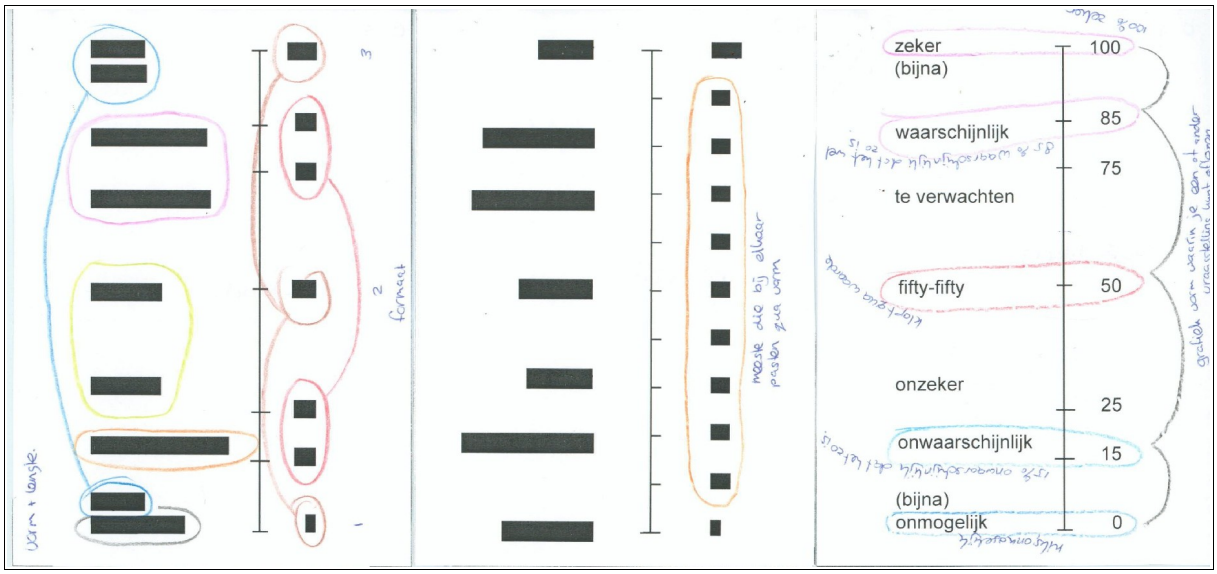


Image 8.4. The sheets of Subject 2.

8.4. Conclusion

We started this experiment because we wanted to know what the cause is of the different behavior we observed in Experiment 1 of the subjects on the two scales. The results from Experiment 2 do not clarify what changes in the original scale were the cause of the different behavior in Experiment 1. The results of comparing the Variable Organization do however confirm that the two representations (organizations) of the scale attract different behavior from its users. The original organization (O1) showed dramatically less horizontal clusters than the alternative organization (O2). Furthermore, in the original organization a higher percentage of form motivations could be observed. And on the alternative scale we observed a higher percentage of positional motivations for clustering.

Another clear result from this experiment was the finding that semantics seem to be a stronger motivation for clustering objects than form and position. The possibility to motivate clusters with form or position seems to be pushed to the background when semantic objects are present in a scale image. Furthermore is seemed that the awareness of differences in form and position of the objects encourages subjects to motivate their clusters with form an position.

We should however not forget that this was a Pilot Experiment. To be more certain about the results this experiment should be repeated in a large quantitative setting and with sub questions instead of one research question.

Chapter 9: Conclusion, recommendations and further research

This chapter contains a summary of the problem descriptions and results of the two experiments followed by the conclusion. Furthermore we give recommendations for better alignment between human experts and probability elicitation methods like the verbal-numerical scale in Section 9.3. We can think of many more studies and experiments concerning the verbal-numerical scale. These ideas for further research are listed in Section 9.4.

9.1. Hypotheses and results

This section starts with a summary of the problem analysis and the hypotheses followed by the assessment of the hypotheses with the results of Experiment 1. In Section 9.1.2 we describe how far we recognized the observations described in Chapter 3 in our results from Experiment 1.

In Chapter 3 we described some observations concerning the use of the verbal-numerical scale in unsupervised group sessions in the EPIZONE project. The elicitor that attended the sessions observed frustration with some subjects about answering the probability questions on the verbal-numerical scale. She also experienced that experts indicated their own certainty about their answer on the scale instead of the answer itself. By analyzing the answer sheets from the EPIZONE sessions we observed that many experts indicated their answer on the scale in another manner than was instructed, which we called wrongly indicated answers. Furthermore it appeared that more than half of all answers had been indicated on an anchor. We stated that these observations lead to problems with the verbal-numerical scale, namely the interpretation and the reliability of the answers. For Experiment 1 we hypothesized about the causes underlying the observed behavior in the EPIZONE project.

9.1.1. Hypotheses and results

Four hypotheses were made concerning the representation of the verbal-numerical scale and three concerning the method.

Representation

We described the representation of the verbal-numerical scale as the position of all objects of the scale in relation to each other, the interval of the labels, the semantics of the labels and the affordance of the scale. Below we list the four hypotheses on the representation followed by the assessment of the hypotheses.

Hypothesis 1: *The affordance of the original scale does not attract users to indicate their answer as a point on the vertical line*

Hypothesis 2: *The affordance of the original scale does attract users to indicate their answer on one of the available anchors.*

Hypothesis 3: *The alternative representation of the verbal-numerical scale results in less wrongly indicated answers and answers indicated on anchors than the original representation.*

Hypothesis 4: *The semantics of the verbal labels 'certain (almost)' and 'uncertain' causes users to indicate how certain they are themselves on the scale instead of indicating the answer to the question.*

Hypotheses 1, 2 and 3 can be confirmed, but hypothesis 4 remains uncertain. The affordance of the original verbal-numerical scale does not attract users to indicated their answer as a point on the vertical line of the scale (hypothesis 1), but does attract users to indicate their answers on one of the available anchors (hypothesis 2). This result is shown in Table 5.5 in which is shown that none of the subjects using the original scale indicated his answer as a point on the vertical line. Instead they indicated 16 out of 17 answers by encircling a numerical anchor. The subjects using the alternative scale indicated more than half of their answers as a point on the vertical line and indicated only 3 out of 16 answers on an anchor. These results do also confirm hypothesis 3. The intention of the designers of the verbal-numerical scale is that subjects indicate an answer on the scale by indicating a point estimate on the vertical line. All other types of answers we called wrongly indicated answers in this thesis. The original scale resulted in twice as much wrongly indicated answers than the alternative scale. Hypothesis 4 cannot be confirmed. It cannot be dismissed either. Five of the six subjects started Part 2 by indicating how certain they were about their answer, instead of indicating the answer itself on the scale. We have no indications that lead us to believe that this was caused by the verbal labels 'certain (almost)' and 'uncertain'. We believe this behavior was caused by the setup of the experiment by letting subjects answer the probability questions first as open answer and secondly on the scale. It is possible that the combination of the setup and the presence of the two verbal labels under consideration caused the observed behavior.

Method

We recall that in Section 3.4.3. we stated that the method is: providing experts with a scale that combines probability words and numbers as a means to support experts in providing a point estimate. It is about the task of providing a point estimate and the use of the concept of probability in the question and answer. Below we list the three hypotheses on the method followed by an assessment of these hypotheses.

Hypothesis 5: *When asked for a probability assessment, users provide an answer in a format different from a point estimate.*

Hypothesis 6: *The combination of probability words with probability numbers in the verbal-numerical scale causes confusion about the use of the scale.*

Hypothesis 7: *Users adjust their answer to fit an existing anchor.*

Hypothesis 5 is partially confirmed. One subject did provide a point estimate when asked for probability assessment. But the other five subjects provided ranges and verbal open answers. Hypothesis 6 cannot be confirmed, but cannot be dismissed either. We observed that at least one subject had trouble interpreting the scale and indicating his answer because of the combination of verbal and numerical labels on the scale. The other subjects had important remarks about the combination of the specific words and numbers, but had no problems with the concept of combining words and numbers (the method) in general. One subject expressed the he thought the combination of words and number could support the users of the scale. Hypothesis 7 is only confirmed for the subjects who used the original scale. Those subjects all adjusted their open answers in Part 3 towards the anchors on the original scale. The subjects who used the alternative scale did not adjust their open answers towards anchors. This seems to be a representational issue more than a method issue. Apparently people are prone to be drawn towards anchor points, but in Experiment 1 it is only the original representation of the scale that seems to attract this behavior from the subjects.

9.1.2. Experiment 1 and the observations

In Experiment 1 we recognized the observations from the EPIZONE project described in Chapter 3. Subjects got frustrated because of probability questions that were not clear enough and because they had the feeling they could not indicate their (open) answer on the original verbal-numerical scale. We did observe that subjects indicated how certain they were about their answer on both representations of the scales. We believe this was due to the setup of the experiment in which the subjects had to answer the same questions twice. All answers on the original scale were wrongly indicated, in contrast with the alternative scale on which only half of all questions was wrongly indicated. Of course, the subjects in Experiment 1 were not instructed how the indicate an answer. It does however show that the representation of the original verbal-numerical scale does not attract intuitive good use of it (affordance). It was observed that five subjects did not formulate their open answer to the probability questions as a point estimate. However, we do not have indications that the format of the initial open answer contributed to indicating answers wrongly. The concept of combining probability words with probability numbers on a scale does not seem to be leading to problems. The specific probability words that were used in the scale do cause some confusions with the interpretation of the scale. Almost all answers on the original verbal-numerical scale were indicated on an anchor, whereas just three answers were indicated on an anchor on the alternative scale. The observation in the EPIZONE project of probabilities not adding up to one could not be observed in Experiment 1, because the subjects did not assess probability questions from the same probability distribution; the probability some symptom is observed and the probability that the same symptom will not be observed under the same circumstances. We suspected in Chapter 3 that subjects might not understand the concept of probability well enough. Experiment 1 showed no indication of problems with the concept of probability. We did find that subject have problems interpreting the probability questions. This is however not caused by the use of probabilities. The

definitions of the terms that were used in the questions caused some discussion and problems with understanding the questions.

9.2. Conclusion Experiment 1 and 2

Overall we conclude that the representation of the original verbal-numerical scale is causing a problem with the reliability of the answers indicated on the scale, because the subjects using this scale adjusted all their open answers towards the seven numerical labels (anchors). The alternative verbal-numerical scale seem to prevent for this problem, because this behavior was not observed with the subjects using the alternative scale. Several changes were made at once to the original scale, therefore the individual influence of the changes cannot be traced. Experiment 2 was setup as a means to investigate what changes to the original scale is causing the subjects to use it differently.

The results from Experiment 2 do not clarify what changes in the original scale were the cause of the different behavior that was observed in Experiment 1. The results of comparing the Variable Organization do however confirm that the two representations (organizations) of the scale attract different behavior from its users. The original organization (O1) showed dramatically less horizontal clusters than the alternative organization (O2). Furthermore, in the original organization a higher percentage of form motivations was observed. And on the alternative scale we observed a higher percentage of positional motivations for clustering.

Another clear result from Experiment 2 was the finding that semantics seem to be a stronger motivation for clustering objects than form and position. The possibility to motivate clusters with form or position seems to be pushed to the background when semantic objects are present in a scale image. With the changes to the representation of the verbal-numerical scale, the semantics were also changed. The modifiers '(almost)' are absent in the alternative representation, but the semantics of the numerical labels is also changed by adding numerical labels and changing numerical labels. Since the semantics seem to be such an important motivation to cluster objects and clustering is a reflection of how something is interpreted, the changed numerical labels in the alternative scale could be preventing for the problems we encountered with the original scale.

The method of providing experts with a scale that combines probability words and numbers as a means to support experts in providing a point estimate does not seem to cause problems. Only one subject had some trouble to interpret the scale because of the combination of words an numbers. He managed however to interpret the question without much help from the elicitor. The use of specific probability words does cause confusion and was criticized by the subjects. The way the probability questions were formulated caused problems with the interpretation of the questions. This was caused by the terms that were used and not by the use of the concept of probability.

9.3. Recommendations

In the introduction we stated that the goal of this thesis is to provide recommendations for better alignment between human experts and probability elicitation methods like the verbal-numerical scale.

Adapting to information need

In Experiment 1, in many cases we had to give feedback to the subjects about the terms that were used in the question and interrupt to prevent that a questions would be wrongly interpreted. Three subjects said they would have liked more information about the terms that were used in the probability questions. We would however not recommend to place large amounts of information next to the probability questions. We think there should be more information about the question on demand of the users. If they want more information it should be available. They need to know that it is available, but it should not be displayed always.

Feedback about interpretation

Most subjects were surprised that their answer on the verbal-numerical scale was interpreted by measuring the distance between the answer and '0' with a ruler. In Experiment 1 it was not told beforehand how we would do that. In the EPIZONE project the subjects were told and shown how their answers would be interpreted before the group session started. We recommend however to give immediate feedback, for every question, of how the answer is interpreted. From Experiment 1 we observed that the subjects were afraid their answer would be used in another way than they intended with their answer. Immediate feedback would give the subject the opportunity to adjust his answer when he sees his answer is interpreted differently than he thought. This will give the subjects the feeling that they are in control of the answers.

Probability words

The probability words that are used in the verbal-numerical scale were criticized by the subjects in Experiment 1 and 2. One subject expected the words 'certain' and 'uncertain' to be opposites, which they are not in the scale. In Experiment 2 a subject clustered the labels 'fifty-fifty', 'uncertain' and '50' with the motivation that fifty-fifty means that you are uncertain. This shows that subject connect the word 'uncertain' to themselves, which is not the intention. Furthermore it shows that this subject thinks 'fifty-fifty' and 'uncertain' have the same numerical value, namely '50'. Another subject would switch the position of the words 'expected' and 'probable'. We would recommend to replace at least the words 'certain' and 'uncertain' by other probability words, because these words really caused confusion about the scale. A possible alternative for the probability words is listed below in Example 9.1. This alternative is not investigated but is proposed because:

- the order of the probability expressions is clear
- the probability expressions fit the kind of expressions Dutch experts use to express their answer to probability questions
- the expressions are a semantic correct answer to the Dutch question “hoe groot is de kans dat (…)” (In English: how high is the chance that (…))
- the expressions can only be interpreted in one way

Example 9.1. (In Dutch):
Hoe groot is dat kans dat dit varken een verminderde eetlust heeft?

Die kans is:

- honderd procent
- heel groot
- redelijk groot
- fifty-fifty
- redelijk klein
- heel klein
- nul

In [4] the designers of the verbal-numerical scale state that they believe that “expressions with modifiers may give more rise to ambiguity than one-word expressions”. We think the modifiers give the opportunity to use probability expressions in the scale that are semantic correct answers to the probability question.

Probability numbers
We showed in Experiment 1 that the original scale was used a seven-point scale, whereas the alternative scale was used as a continuous scale on which any degree of probability could be indicated. We believe that the the changes made to the numerical labels could be the cause of this. In the original scale the right column was designed as a scale with ten numerical labels with equal intervals, opposite poles and a neutral midpoint. Therefore we recommend to use the right column of the alternative scale for the verbal-numerical scale.

9.4. Further research

The verbal-numerical scale as use for probability elicitation from experts turned out to be an interesting and complex domain which can be explored much further than was done in this thesis.

As an alternative of supervising the probability elicitation sessions the possibility of eliciting probabilities in a digital environment should be investigated. We determined that some feedback during the probability elicitation with the verbal-numerical scale is necessary. A digital application could provide that feedback. Furthermore, with an application it would be easier and faster to analyze the results of the sessions. The application could provide some feedback like drawing the subjects attention to chances that should add up to one. It would also be possible to give more information about the terms that are used in the probability question if the subject want more information, for instance by hovering with the mouse over a term in the question. It could also be worthwhile to experiment with tools to indicate the answer on the scale to prevent for wrongly indicated answers. The use of a slider indicator would be interesting to investigate (see example in Image 9.1.).

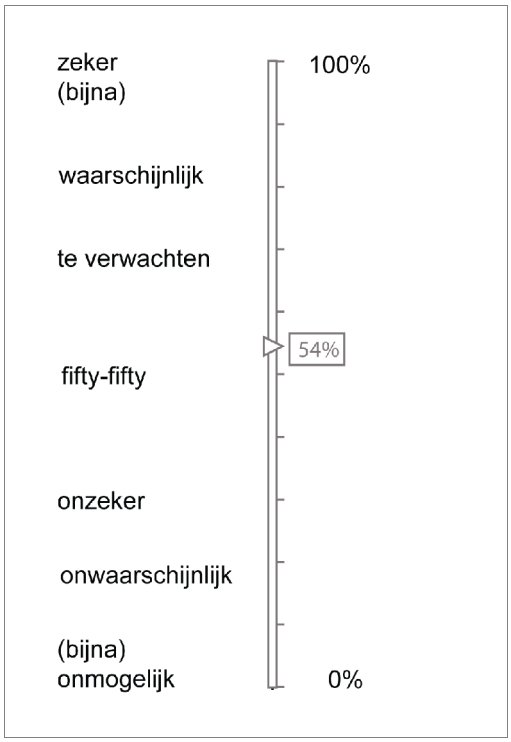


Image 9.1. Example of digital verbal-numerical scale with slider indicator.

We believe that the problems with the verbal-numerical scale in unsupervised sessions could be solved when a digital application for the sessions will be carefully designed.

References

[1] Fred Hamburg (1999). *Elementaire Besliskunde. Over de verwevenheid van ratio, emotie en intuïtie.* Utrecht: Uitgeverij LEMMA BV

[2] Vicky Sauter (1997). *Decision Support Systems: An Applied Approach.* New York, John Wiley.

[3] Renooij (2001*). Probability elicitation for belief networks: issues to consider, issues to consider and existing methods for probability elicitation. The Knowledge Engineering Review*, Vol. 16:3, 255–269 (ook in chapter 2)

[4] Silja Renooij, Cilia Witteman (1999). *Talking probabilities: communicating probabilistic information with words and numbers*. International Journal of Approximate Reasoning, 22, 169-194. (ook chapter 2)

[5] L.C. van der Gaag, S. Renooij, C.L.M. Witteman, B.M.P. Aaleman, B.G. Taal (2002). *Probabilities for a probabilistic network: A case-study in oesophageal cancer.* Artificial Intelligence in Medicine, vol. 25 (2), pp. 123-148.

[6] P.L. Geenen, A.R.W. Elbers, L.C. van der Gaag, W.L.A. Loeffen (2006). *Development of a probabilistic network for clinical detection of classical swine fever.* Proceedings of the 11th Symposium of the International Society for Veterinary Epidemiology and Economics, Cairns, Australia, pp. 667669.

[7] Charitos Th., Van der Gaag L.C., Visscher S., Schurink K., Lucas P.J.F. (2005). *A dynamic Bayesian network for diagnosing ventilator-associated pneumonia in ICU patients.* Proceedings of the 10th Intelligent Data Analysis in Medicine and Pharmacology Workshop 2005:32-37.

[8] CSF EU bijeenkomst, provided by Silja Renooij by mail on 17th December 2008.

[9] Linda van der Gaag (2008). Personal conversation

[10] Michael Negnivitsky (2005). *Artificial Intelligence. A guide to Intelligent Systems.* Second edition. Harlow: Pearson Education.

[11] Daniel Kahneman, Paul Slovic, Amos Tversky (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press

[12] Wallsten TS, Budescu DV, Zwick R. (1993). *Comparing the calibration and coherence of numerical and verbal probability judgments*. Manage Sci;39:176–90.

[13] Seymour Sudman, Norman M. Bradburn, Norbert Schwarz (1996). *Thinking About Answers. The Application of Cognitive Processes to Survey Mthodology.* San Francisco: Jossey-Bass Publishers

[14] Donald A. Norman (1999). *Affordances, Conventions and Design*. Interactions 6(3):38-43. ACM Press.

[15] Rudolf Arnheim (1969). *Visual thinking*. California: University of California Press

[16] Charles Kostelnick, David D. Roberts (1998). *Designing Visual Language. Strategies for professional communicators*. Needham: Allyn & Bacon.

[17] Mosteller F., Youtz C. (1990). Quantifying probabilistic expressions, Statistical Science 5 2-12.

[18] Ericsson, K.A., & Simon, H.A. (1993). *Protocol analysis: verbal reports as data (Rev. Ed.).* Cambridge, MA: The MIT Press.

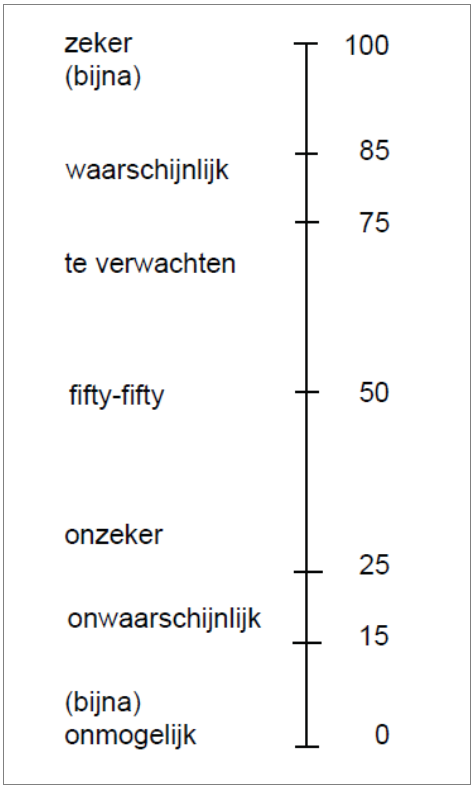
[19] John R. Anderson (2000). *Cognitive Psychology and its implications*. Fifty Edition. New York: Worth Publishers. (p. 47-49)

[20] *The legacy of max Wertheimer and gestalt psychology.* Magazine: Social Research, winter, 1994.

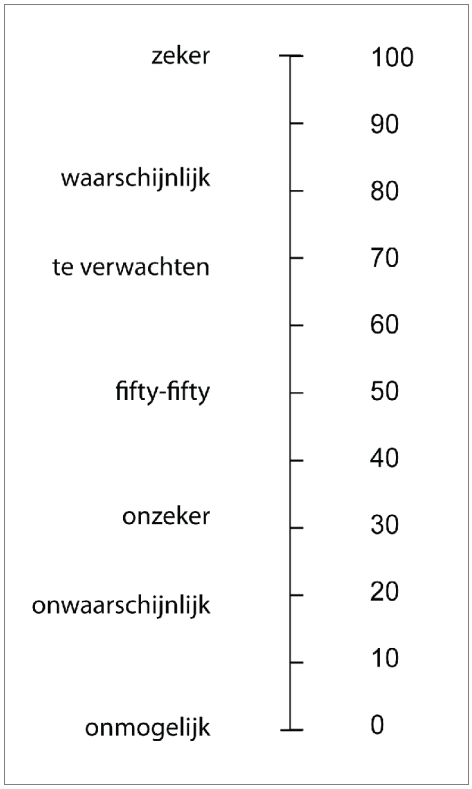
[21] Irvin Rock and Stephen Palmer (1990). *The legacy of Gestalt Psychology.* Scientific American December 1990

[22] *Art, Design and Gestalt theory*, Magazine Leonardo, 1998, section Historical perspective.

Appendix A: Verbal-numerical scale in Dutch



Original representation



Alternative representation

Appendix B: Probability Vignettes in Dutch

Oefenvraag

Stelt u zich een willekeurig *varken* voor van een willekeurig bedrijf. Hoe waarschijnlijk is het dat dit varken in de afgelopen 2 weken *curatief* is *behandeld met antibiotica*?

Cyanose vraag: (2 kenmerken, kans 0.33)

Stelt u zich een *varken* voor met een *circulatiestoornis*. Hoe waarschijnlijk is het dat dit varken *cyanose* (blauw/paarsverkleuring) van de oren en/of andere lichaamsuiteinden (poten, neus, rond de staart) vertoont?

Pr(Cyanosis = ja | Circulatiestoornis = ja) = 0.33

Extra informatie:

We hebben het hier dus eigenlijk over zowel perifere cyanose (tgv shock: bloeding uit vaten) als over centrale cyanose (tgv luchtweginfectie -> ademhalingsproblemen). De circuclatiestoornis kan daarbij dus of een herverdeling/vertraging van de bloedstroom zijn of een tekort aan zuurstof (wat wellicht ook tot vertraging van de bloedstroom leidt, omdat de hartspier dan op den duur ook te weinig zuurstof krijgt en minder hard gaat pompen. Omdat we geen kansen voor het model aan het schatten zijn, maakt het op zich niet zoveel uit. Je kan dus denk ik wel zeggen dat een circulatiestoornis een vertraagde bloedstroom is die verschillende oorzaken kan hebben, maar die zijn dan niet relevant.

Conjuncivitis vraag: (2 kenmerken, kans 0.85)

Stelt u zich een *varken* voor met een *slijmvliesontsteking* van de voorste luchtwegen (neus/pharynx/larynx). Hoe waarschijnlijk is het dat dit varken een zichtbare *conjunctivitis* (roodheid, traanstrepen) aan één of beide ogen heeft?

Pr(Conjunctivitis = ja | Slijmvliesontsteking = ja) = 0.85

Extra informatie:

De vraag gaat over conjunctivitis, dus over echte, zichtbaar door de aanwezigheid van traanstrepen; het gaat niet over de kans op zichtbare traanstrepen die door van alles en nog wat veroorzaakt zouden kunnen zijn.

Malaise vraag: (3 kenmerken, kans 0.85)

Stelt u zich een *varken* voor dat *algehele malaise* vertoont, maar een *normale lichaamstemperatuur* heeft. Hoe waarschijnlijk is het dat dit varken een *verminderde eetlust* heeft?

Pr(Eetlust = verminderd, Algehele Malaise = ja, Lichaamstemperatuur = normaal) = 0.85

Extra informatie:

Dit varken is geen slijter.

Stapelen vraag: (3 kenmerken, kans 0.66)
Stelt u zich een *gespeende big* voor met *koorts* (boven de 40,5°C). Bovendien neemt u waar dat er een *suboptimaal stalklimaat* (te koud en tochtig) heerst in de stal waarin dit dier zich bevindt. Hoe waarschijnlijk is het dat dit dier *dicht opeengepakt* met de andere biggen ligt ('stapelen')?

Pr('Huddling' = ja | Type varken = gespeende big, Lichaamstemperatuur = verhoogd, Klimaatprobleem = ja) = 0.66

Intra-uteriene infectie vraag: (4 kenmerken, kans 0.65)
Stelt u zich een *zogende zeug* voor die momenteel uitsluitend *geïnfecteerd* is met het varkenspestvirus. Deze zeug heeft een *late intra-uterine infectie* gehad tijdens de dracht (gedurende de laatste 18 dagen). Er werden totaal 13 volgroeide biggen geboren. Hoe waarschijnlijk is het dat er 3 of meer van deze 13 biggen dood werden geboren?

Type Varken = zeug .Pr(Doodgeboren biggen = ja | Reproductie fase = onlangs geworpen (zogend), Intra-uterine infectie = laat, Andere primaire infecties = geen, Viraemie KVP = ja) = 0.65

Reserve vraag: (2 kenmerken, kans 0.35)
Stelt u zich een varken voor met *koorts* (boven de 40,5 °C). Bovendien neemt u waar dat er een *suboptimaal stalklimaat* (te koud en tochtig) heerst in de stal waarin dit dier zich bevindt. Hoe waarschijnlijk is het dat dit dier *lange haren* toont (de haren staan zichtbaar overeind)?

Pr(Haren overeind = ja | Lichaamstemperatuur = verhoogd, Klimaatprobleem = ja) = 0.35

Appendix C: Practice Questions Experiment 1 in Dutch

Som - vermenigvuldigen

34
27 x

Anagram
kinewl

Anagram
afotobulm

Anagram
grlichtvein

Oefenvraag
Stelt u zich een willekeurig *varken* voor van een willekeurig bedrijf. Hoe waarschijnlijk is het dat dit varken in de afgelopen 2 weken *curatief* is *behandeld met antibiotica*?

Geef hierop het antwoord dat u aan een collega-dierenarts zou geven:

Appendix D: Protocol Experiment 1 in Dutch

Protocol

Binnenkomst

[Voorstellen, praatje maken, etc.]
Ik ga van nu af aan alles voorlezen, want het is heel belangrijk dat ik bij alle proefpersonen hetzelfde zeg.

Wat gaan we doen

Wat we straks gaan doen is een zogenaamd thinkaloud experiment, een experiment waarbij u hardop zegt wat u denkt. Straks zal ik, als u dat wil, meer uitleggen over mijn project en kunt u alles vragen.

Video en audio

U ziet al dat ik een camera en een microfoontje meegenomen heb. Zoals u zich wellicht kunt voorstellen kan ik niet precies onthouden wat u gezegd heeft in dit experiment, dus zal ik het opnemen. De videocamera is alleen op het papier gericht, niet op uw gezicht en uw lichaam. Ik wil graag benadrukken dat dit experiment anoniem is, u wordt gewoon een nummer en het is nergens af te leiden wie wat heeft gezegd en gedaan.

U gaat straks vijf vragen beantwoorden. Het experiment bestaat uit drie delen. Na afloop is er nog een korte vragenlijst over het experiment en kunt u uw commentaar en suggesties kwijt.

[apparatuur opstellen] *Ik ben wel eventjes bezig om de apparatuur op te stellen, pakt u ondertussen wat te drinken voor uzelf.*

De apparatuur staat nog niet aan.
We doen dit experiment omdat we graag willen weten wat u denkt terwijl u het antwoord op aantal vragen geeft in uw expertgebied, die ik u straks zal voorleggen. Ik ga u daarom vragen om hardop te denken terwijl u de vraag beantwoordt. Met hardop denken bedoel ik dat u alles zegt wat u denkt, vanaf het moment dat u de vraag ziet, totdat u een antwoord gegeven heeft. Ik wil dat u voortdurend hardop praat vanaf dat u de vraag ziet totdat u uw uiteindelijke antwoord gegeven heeft. Het is niet de bedoeling dat u van tevoren nadenkt over wat u gaat zeggen, of dat u aan mij probeert uit te leggen wat u zegt. Stelt u zich voor dat u alleen in deze ruimte bent en tegen uzelf spreekt. Het belangrijkste is dat u blijft praten. Als u een tijd stil bent, zal ik u vragen om door te praten.

Begrijpt u wat de bedoeling is?

Ik wil benadrukken dat wij geïnteresseerd zijn in de manier waarop u de vragen beantwoord en hoe u uw antwoord op papier zet. Het gaat ons niet om het antwoord zelf. U kunt een vraag niet fout

beantwoorden, alle antwoorden zijn goed. Wel wil ik dat u uw best doet om de vragen zo goed mogelijk te beantwoorden.

Goed, we zullen beginnen met een aantal oefenvragen. Het is heel eenvoudig.

Ik zet de camera en het microfoontje alvast aan, om te kijken of alles goed werkt.
[voorbeeldsom zelf voordoen zoals het wel en zoals het niet moet]
[pp doet som/anagram] [feedback geven]

Oefenvragen

42
25 x

Als de pp er niet uit komt: *het gaat om uw denkproces.*
En dit is een moeilijk anagram, dus er moet even veel gedacht worden, dan kan ik goed zien of u het think aloud goed doet. Het gaat er helemaal niet om of u het kunt oplossen! De meeste mensen komen er niet uit . . .

Oefenvragen

[pp oefenvraag geven]

Begint u met het voorlezen van de vraag en probeert u daarna door te praten. Als u wilt mag u aantekeningen maken, maar dat hoeft niet. Graag in ieder geval uw antwoord op het papier zetten.

[pp voert oefenvraag uit]
[pp helpen met praten door neutrale vragen te stellen of pp aan te moedigen. Bijvoorbeeld: 'en wat denk u nu?' en 'probeert u door te praten'.]
[reactie geven op uitvoering, afhankelijk van situatie]

Dat ging heel goed.
Hoe vond u het om hardop te praten?
Een beetje onwennig hè in het begin? Dat heeft iedereen wel hoor.

[checken of de apparatuur het goed gedaan heeft. Eventueel laten zien wat is opgenomen (alleen de handen en het papier). Apparatuur weer aanzetten.]

Aan de slag

Zo, dan gaan we nu met het experiment beginnen. Als u het antwoord op een vraag heeft opgeschreven, slaat u het blaadje om en gaat u door met de volgende vraag.
Heb ik het zo duidelijk uitgelegd? Heeft u nog vragen?
Goed, dan gaan we beginnen met het eerste deel.

Experiment deel 1

[geef vraagpakketje pp]

[als er als open antwoord puntschattingen zijn gegeven, dan van die schattingen vragen of ze inderdaad precies dat antwoord bedoelen]

Zullen we direct doorgaan met het tweede deel of wilt u liever even pauzeren?

Experiment deel 2

Ik ga u vragen uw eerder opgeschreven antwoord steeds per vraag op een antwoordschaal weer te geven. Ook hierbij is het de bedoeling dat u weer hardop denkt.
Begint u met het voorlezen van de vraag, en daarna het antwoord dat u gegeven hebt, probeert u daarna door te praten als u uw antwoord op de antwoordschaal gaat weergeven. Heb ik het zo duidelijk uitgelegd? Heeft u nog vragen?

Neemt u hier rustig de tijd voor.

[eventuele aanvullende instructie] Dus probeert u de vraag nogmaals te beantwoorden, maar zet dan uw antwoord op de antwoordschaal. Begint u weer met het voorlezen van de vraag en probeert u dan door te praten terwijl u uw antwoord op de antwoordschaal zet.

[eventueel aanvullende instructie geven] *Het is de bedoeling dat u dit* [aanwijzen] *antwoord op deze* [aanwijzen] *antwoordschaal weergeeft.*

[Als pp de schaal niet goed gebruikt wordt pp, nadat alle vragen zijn beantwoord, aangemoedigd om het antwoord op een andere manier als 1 punt op de schaal te zetten. Dit gebeurt op een schoon blaadje met de schaal.]

Experiment deel 3

We gaan nu terugluisteren hoe u de vragen heeft beantwoord. Na elke vraag zet ik het geluid even stop en vraag ik u mij alles te vertellen dat u zich kunt herinneren over uw gedachten terwijl u de vraag beantwoordde.

Experiment deel 4

Oké, heel goed. Dan gaan we nu door naar het laatste onderdeel van het experiment.

Over het algemeen worden deze kansen door een AIO verzameld, een assistent in opleiding. Iemand anders interpreteert later het antwoord. Diegene ziet alleen de vraag en uw antwoord op de antwoordschaal en dus niet uw redeneringen. Ik zal u laten zien hoe diegene uw antwoord interpreteert. Graag hoor ik uw reactie op de interpretatie.

[Per vraag met liniaal percentage aflezen en getal ernaast zetten]

Wat u hier aangeeft wordt geïnterpreteerd als precies 20%. Klopt dit? [zo nee] Welke interpretatie had u liever gewild?

Had u verwacht dat de antwoordschaal zo geïnterpreteerd zou worden? [zo nee] *Hoe had u dan verwacht dat de antwoordschaal geïnterpreteerd zou worden?*

Oké, het experiment is nu klaar, u hoeft niet meer hardop te denken. Ik stel u nog een aantal vragen over het experiment.

- *Vond u de vragen duidelijk?*
- *Hoe vond u het om uw (open) antwoord te geven?*
- *Hoe vond u het om uw (open) antwoord op de antwoordschaal te zetten?*
- *Wat vindt u van de combinatie van woorden en getallen op de antwoordschaal?*
- [bij woorden] *hoe zou u dit als punt aangeven, [bij getal] hoe zou u dit in woorden omschrijven?*
- *Wat vond u van het experiment?*
- *Heeft u verder nog commentaar/advies enz.*

Oké, ik ga de apparatuur nu uitzetten en dan wil ik u vragen om ondertussen een kort vragenlijstje in te vullen.

Questionnaire

[geef questionnaire aan pp, afwenden, niet praten, pp zijn gang laten gaan en apparatuur uitzetten]

Toestemmingsformulier

Ik wil u hartelijk bedanken voor uw medewerking. Tot slot heb ik hier voor u op papier nog eens een uitleg van mijn onderzoek en wat ik met de resultaten uit het experiment ga doen.
Wilt u dat ik het aan u uitleg of wilt u het straks zelf lezen?
U kunt op het formulier aangeven of u na afloop een kopie van mijn scriptie wil ontvangen.
Ik wil u vragen niets over dit experiment te bespreken met andere dierenartsen.
[geef toestemmingsformulier]

Afronden

Heeft u nog vragen of opmerkingen over mijn onderzoek of over dit experiment?
[opmerkingen eventueel noteren]

Dan wil ik u graag hartelijk bedanken voor uw tijd. U heeft mij erg geholpen.

Appendix E: Consent Form in Dutch

Afstudeerproject Gwyneth Ouwehand

Afstudeerproject

De titel van mijn afstudeerproject is:

"Supporting computer-aided decision making by improving alignment between human experts and probability elicitation instruments."

Ik zal dit hieronder proberen duidelijk toe te lichten.
We hebben het niet altijd in de gaten, maar we gebruiken de hele dag door kansen. We schatten de kans in dat het gaat regenen als we op pad gaan, de kans dat we promotie krijgen op het werk en de kans dat we winnen in de loterij.
Kansen worden ook gebruikt om systemen te ontwikkelen die mensen kunnen helpen in het nemen van beslissingen. Beslissingsondersteunende systemen noemen we dat.
Om zo'n systeem te maken wordt een netwerk ontwikkeld van een domein waarin alle variabelen staan en welke variabelen invloed op elkaar hebben. Voor dit netwerk zijn ook kansen nodig; de kans dat als het ene waar is, iets anders ook waar is.
De informatie die nodig is voor zo'n systeem wordt geleverd door zogenaamde experts, ervaren mensen in het vakgebied. Het valt echter niet mee om kansen te schatten. Wij mensen zijn daar nou eenmaal niet goed in.
Probeert u maar eens te zeggen wat de kans is dat het morgen gaat regenen. Hoogstwaarschijnlijk zullen de meeste mensen een globaal beschrijvend antwoord geven, zoals 'best groot'. Voor het netwerk is het echter van belang of dat 'best groot' gelijk staat aan 80% kans op regen of 85% kans op regen.
Het gat tussen de natuurlijke kansbeschrijving en de (exacte) kansbeschrijving die nodig is voor zo'n netwerk willen we graag overbruggen.

Door te onderzoeken wat de natuurlijke manier is van experts om antwoord te geven op kansvragen, kunnen we wellicht een manier bedenken om dit gat kleiner te maken en het verzamelen van kansen voor het netwerk gemakkelijker.

Wat wordt met de uitkomst van dit experiment gedaan?

In het experiment zijn de volgende gegevens van belang:

- Uw open antwoorden; ik wil bekijken in wat voor vorm u van nature antwoord geeft op kansvragen.
- Uw weergave van de antwoorden op de schaal; ik wil zien of u de schaal goed gebruikt heeft en of het antwoord dat u op de schaal geeft overeen komt met uw open antwoord.

- Uw gedachtegang, zoals vastgelegd tijdens het hardop denken; hieruit wil ik proberen te begrijpen hoe u tot een antwoord komt, hoe u de schaal interpreteert en hoe u uw open antwoord als het ware vertaalt naar een antwoord op de schaal.
- Uw antwoorden op de vragen; andere proefpersonen hebben een andere antwoordschaal gezien dan u, ik ben benieuwd of de proefpersonen de ene schaal duidelijker vonden dan de andere schaal.

Anonimiteit

Uw bijdrage aan dit experiment is anoniem. Uw antwoorden worden anoniem verwerkt en uw naam zal niet worden genoemd in de resultaten.

Toestemming

Ik vraag uw toestemming om de audio en video van dit experiment (indien nodig) te presenteren ter ondersteuning van mijn resultaten. Hierbij wil ik nog eens benadrukken dat u nooit herkenbaar in beeld wordt gebracht; wel zou iemand uw stem kunnen herkennen, al is die kans zeer klein.

- ☐ Ik vind het ok als de beelden/audio van dit gesprek gebruikt worden ter ondersteuning van het presenteren van resultaten.
- ☐ Het bovenstaande, maar alleen nadat ik de beelden/audio heb gezien/gehoord, en mijn toestemming heb gegeven.
- ☐ Het bovenstaande nooit.

Desgewenst aankruisen:

- ☐ Na afronding ontvang ik graag een kopie van de scriptie.

Naam: [van tevoren invullen]

Datum:.. …/……/…… [van tevoren invullen]

Handtekening:

Contactgegevens afstudeerder

Gwyneth Ouwehand
De Duikerhof 9
3461 HR Linschoten

Master: Content and Knowledge Engineering
Tel: 0624895547
E-mail: gouwehan@cs.uu.nl
Studentnummer: 0371742

Universiteit Utrecht

Appendix F: Open answers and answers on scale

Subject 1
Q1 (cyanose vraag)
open: <0.01%
schaal: accolade naast de numerical labels van ongeveer 0 – 10. Daarnaast schrijft hij 0-10

Q2 (conjunctivitis vraag)
open: <0.01%
schaal: accolade naast de numerical labels van ongeveer 0 – 10. Daarnaast schrijft hij 0-10

Q3 (malaise vraag)
open: < 5 <0.5
schaal: cirkeltje om getal dertig

Q4 (stapelen vraag)
open: < 0,005%
schaal: accolade naast de numerical labels van ongeveer 0 – 10. Daarnaast schrijft hij 0-10

Q5 (intra-uteriene infectie vraag)
open: 5% - 10%
schaal: cirkeltje om getal twintig

QR (reserve vraag)
open: geringe kans
schaal: cirkeltje om getal 30

Subject 2
Q1 (conjunctivitis vraag)
open: 10 a 20% bij inl. Virale infecties
schaal: cirkeltje om 15 waarbij een stukje van het streepje op die hoogte is meegenomen

Q2 (stapelen vraag)
open: 70 a 80 % kans op huddelen
schaal: cirkel omvat zowel het getal 75 en 85

Q3 (intra-uteriene infectie vraag)
open: 10 a 20 %
schaal: cirkeltje om getal 15 waarbij het streepje op die hoogte en de lijn op die hoogte is meegenomen

Q4 (malaise vraag)
open: 90 a 100% geen of nagenoeg geen voeropname

schaal: grote cirkel om getal 100 waarbij duidelijk de onderkant verder uitschiet onder de 100 dan boven

Q5 (cyanose vraag)
open: 70 a 80 %
schaal: cirkel om getal 75 en het bijbehorende streepje op de verticale lijn

QR (reserve vraag)
open: 90 a 100 % van de dieren
schaal: cirkel om het getal 100 met een grote haal naar beneden waarbij een deel van de verticale lijn wordt meegenomen

Subject 3
Q1 (malaise vraag)
open: 90 %
schaal: kruisje op de verticale lijn ter hoogte van 85. Precies tussen het streepje van de 80 en 90 in.

Q2 (intra-uteriene infectie vraag)
open: 80 %
schaal: kruisje op het streepje op de verticale lijn naast het getal 60

Q3 (conjunctivitis vraag)
open: 70%
schaal: kruisje net onder het streepje naast de 80

Q4 (cyanose vraag)
open: 60%
schaal: kruisje net onder het streepje naast het getal 60

Q5 (stapelen vraag)
open: 100%
schaal: kruisje op het streepje naast het getal 100

QR (reserve vraag)
open:
schaal:

Subject 4
Q1 (stapelen vraag)
open: door tocht en lage temp. zullen bijv altijd al geneigd zijn om 'huddeling' te vertonen, zeker wanneer er ook nog sprake is van hoge koorts.
Schaal: cirkeltje die beide getallen 75 en 80 omvat

Q2 (cyanose vraag)

open: alleen indien sprake is van ernstig hartfalen en/of longaandoening zal een cyanose zichtbaar worden. Deze correlatie hoeft geenszins aanwezig te zijn.
Schaal: cirkeltje om het getal 50

Q3 (malaise vraag)
open: het lijkt me erg waarschijnlijk, dat een dergelijk varken verminderde eetlust heeft, maar geen acute ontstekingen.
Schaal: cirkeltje om getal 85

Q4 (intra-uteriene infectie vraag)
open: dit lijkt me erg ontwaarschijnlijk, omdat dit immuuncompetente biggen zijn en het V.P. Virus niet als lich. Vreemd antigeen zullen beschouwen.
Schaal: cirkeltje om getal 85. Later cirkel om getal 15.
Opmerking: eerste keer per ongeluk zekerheid aangegeven.

Q5 (conjunctivitis vraag)
open: door de conjunctivitis zullen waarschijnlijk traanstrepen ontstaan doordat het traanvocht onvoldoende normaal afgevoerd zal worden vanwege de rhinitis.
Schaal: cirkeltje om getal 75

QR (reserve vraag)
open: de relatie tussen suboptimaal klimaat en lange haren is evident, zeker indien ook nog sprake is van koorts.
Schaal: cirkeltje om getal 85

Subject 5
Q1 (intra-uteriene infectie vraag)
open: 0 < 1% kans
schaal: cirkel om verbal label 'onmogelijk'. Vanuit de cirkel is een lijn met een pijl getrokken naar de vertical lijn net boven het streepje dan bij onmogelijk/0 hoort.

Q2 (stapelen vraag)
open: niet gestapeld maar aan de rand van de groep of solo < 2 %
schaal: cirkel om verbal label 'onmogelijk'. Vanuit de cirkel is een lijn met een pijl getrokken naar de verticale lijn ongeveer op een kwart van de lijn vanaf de 0 naar de 10.

Q3 (conjunctivitis vraag)
open: >80% → 100%
schaal: een accolade-achtige boog tussen de kolom met verbal labels en de verticale lijn. De boog begint ongeveer bij het streepje van de 80 en eindigt ongeveer bij het streepje van de 100. Op het scherpste deel van de boog staat een pijl getrokken. Daarbij staat het woord 'verwachting'.

Q4 (cyanose vraag)
open: niet

schaal: cirkeltje om het verbal label 'onwaarschijnlijk'. Links en rechts van de cirkel is een klein horizontaal streepje gezet.

Q5 (malaise vraag)
open: 95% bijna 100% zeker van verminderde eetlust
schaal: cirkeltje om het woord 'zeker'. Vanuit het cirkeltje is een lijn met een pijl getrokken naar de vertical lijn net net net vlak onder het streepje van 100.

QR (reserve vraag)
open:
schaal:

Subject 6
Q1 (cyanose vraag)
open: cyanose aan oren is kansrijk. Cyanose aan poten/staart alleen bij ernstige stoornis, maar meestal al dood vóór dit extreme symptoom.
Schaal: cirkeltje om het getal 75

Q2 (intra-uteriene infectie vraag)
open: kans is groot wegens kortere incubatiemogelijkheid en sterke gevoeligheid voor KVP-virus van ongeboren vrucht
schaal: cirkeltje om het getal 75

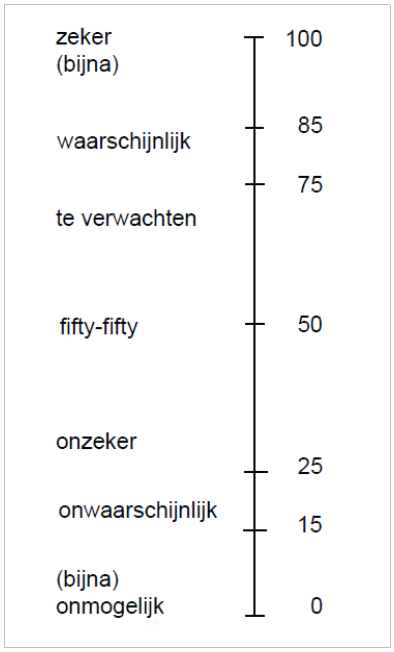
Q3 (stapelen vraag)
open: kans is zeer groot door zoektocht naar warmte (huddeling) KVP pas aan de orde bij specifieke stal/gebiedsverdenking hier wellicht kou + coli/strept. → OK
schaal: cirkeltje om het getal 85

Q4 (malaise vraag)
open: kans is groot (85%). Andere 15% komt voor, maar dan is groei slecht. Malaise door...ziektes eerder bij bv. coli/strept | PRRS of combinatie.
Schaal: cirkeltje om getal 75

Q5 (conjunctivitis vraag)
open: kans is groot (70%). Hierbij zet hij later een lijn naar 50% en naar 85%. Hangt af van combinatie van factoren: NH3 + PRRS + influenza + bezetting + stof.. Dan volgt een streep over de gehele breedte van de bladzijde. KVP pas in geding bij meer specifiek symptomen (petechien, huddling, hoge T, anorexia alom of specifiek.).
Schaal: omcirkeling van het getal 75

QR (reserve vraag)
open: kans niet groot | je ziet het niet. Alleen chronisch gevallen ontwikkelen lange haren en malaise.
Schaal: omcirkeling getal 25.

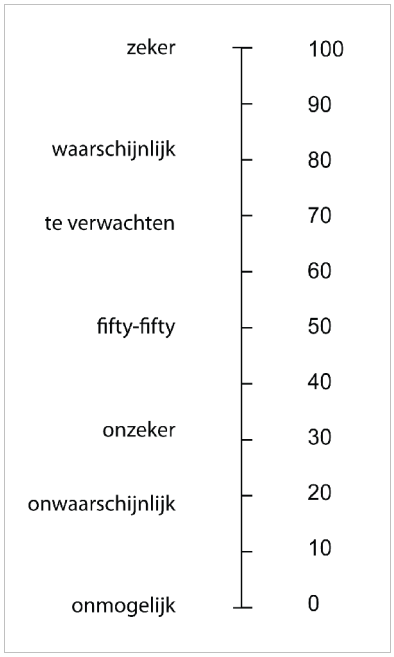
Appendix G: Scale images Experiment 2



O1-S1-R1

O1-S1-R2

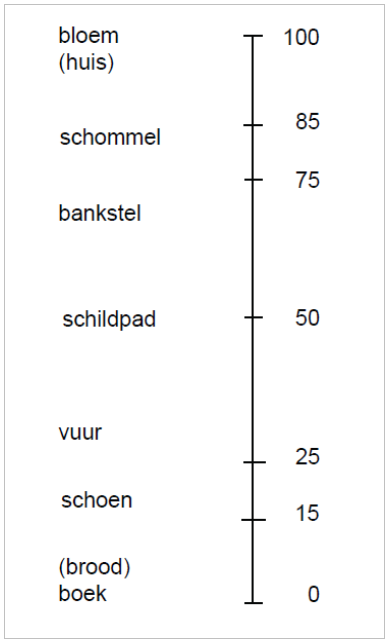
O1-S1-R3



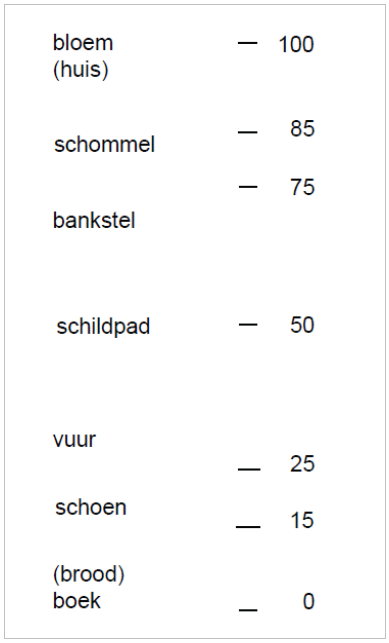
O2-S1-R1

O2-S1-R2

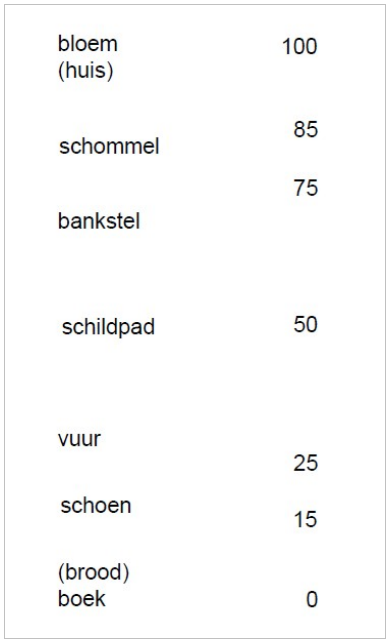
O2-S1-R3



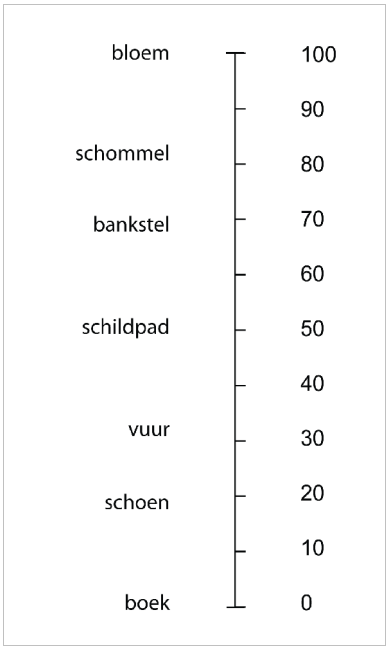
O1-S2-R1



O1-S2-R2



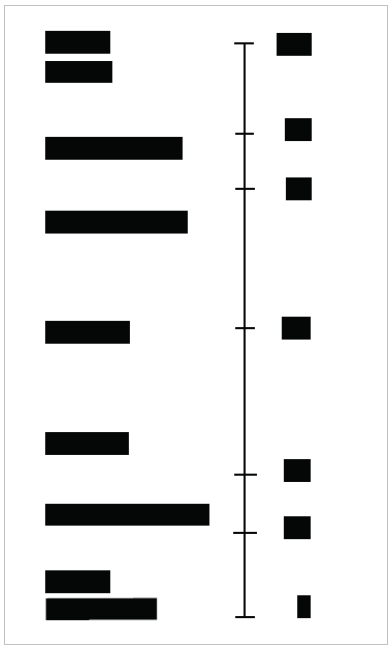
O1-S2-R3



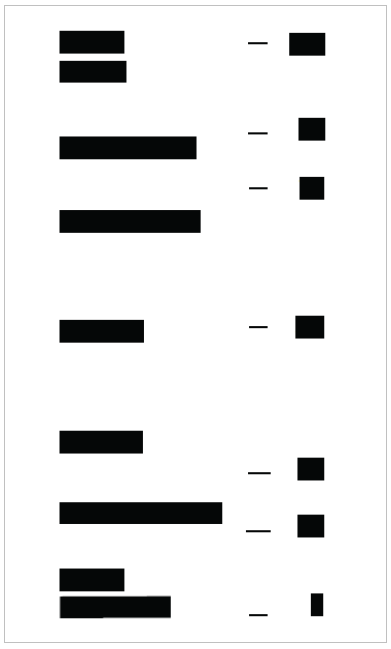
O2-S2-R1

O2-S2-R2

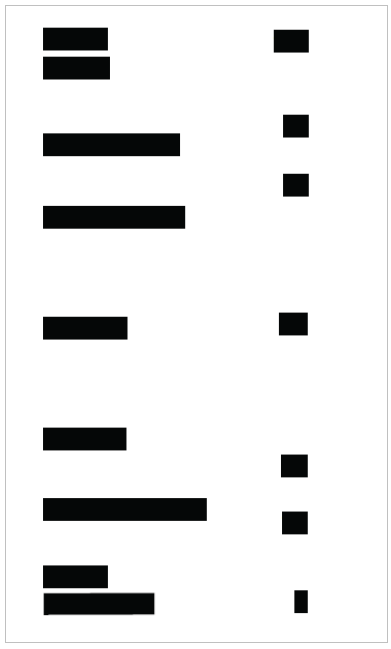
O2-S2-R3



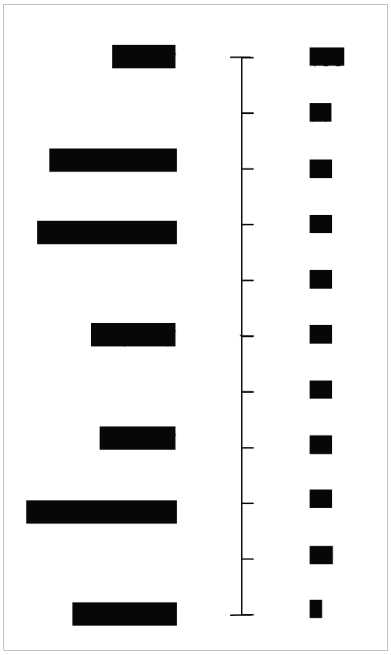
O1-S3-R1



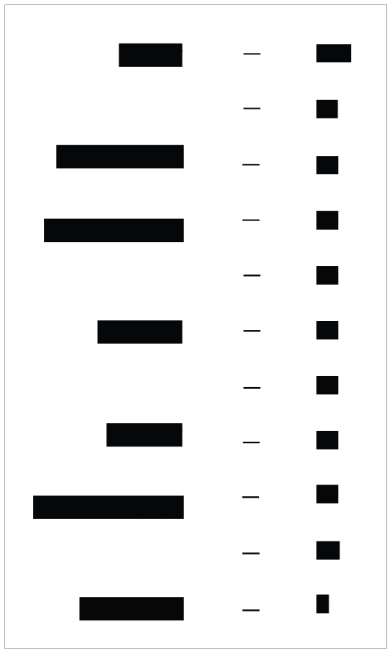
O1-S3-R2



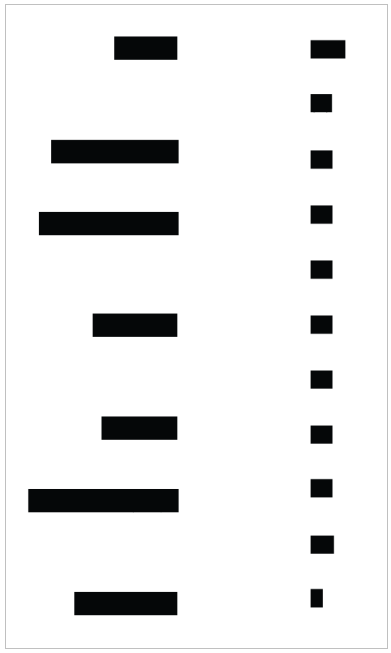
O1-S3-R3



O2-S3-R1



O2-S3-R2



O2-S3-R3

Appendix H: Assignment Experiment 2

Op de volgende pagina's ziet u steeds een afbeelding staan.
Elke afbeelding bestaat uit objecten. Voorbeelden van objecten zijn:

50

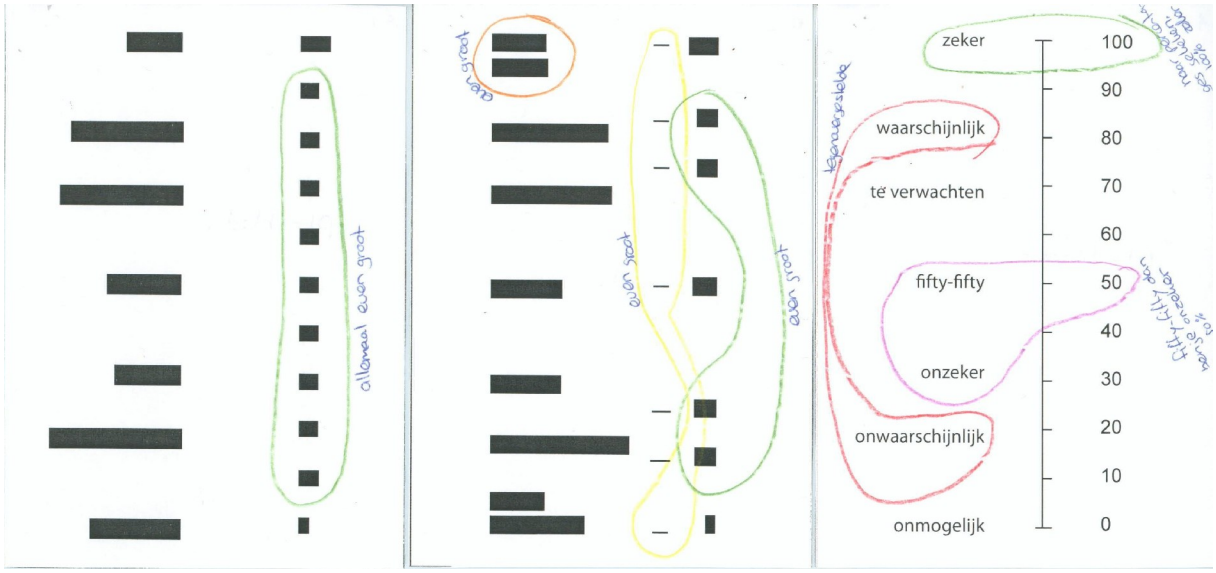
-

|

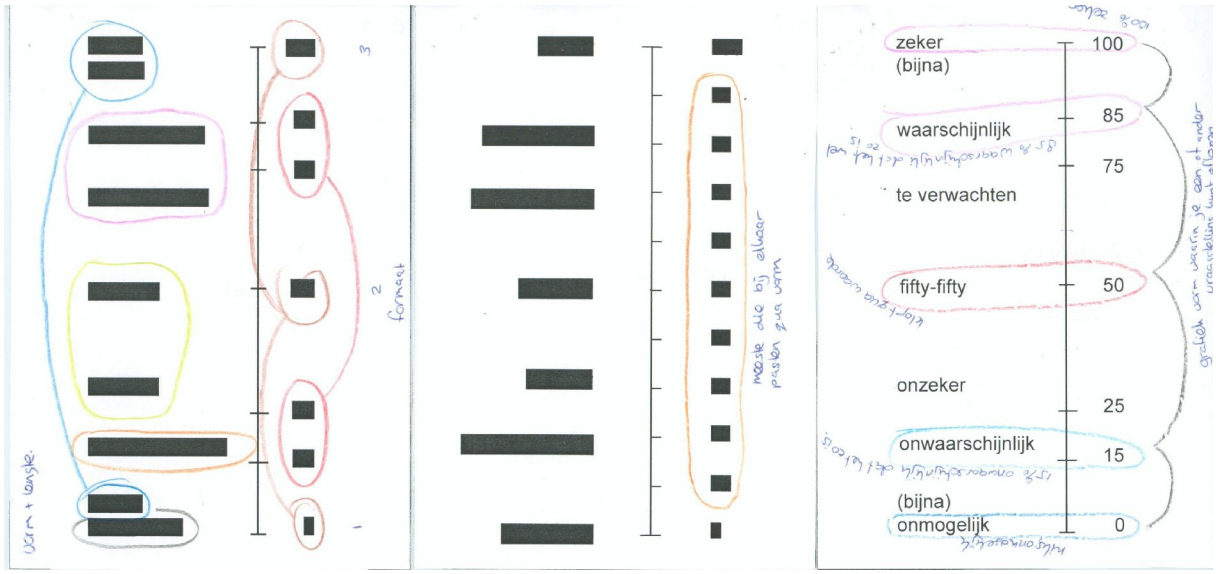
Bloem

Opdracht:
Groepeer de objecten waarvan u denkt dat ze bij elkaar horen door ze te omcirkelen.

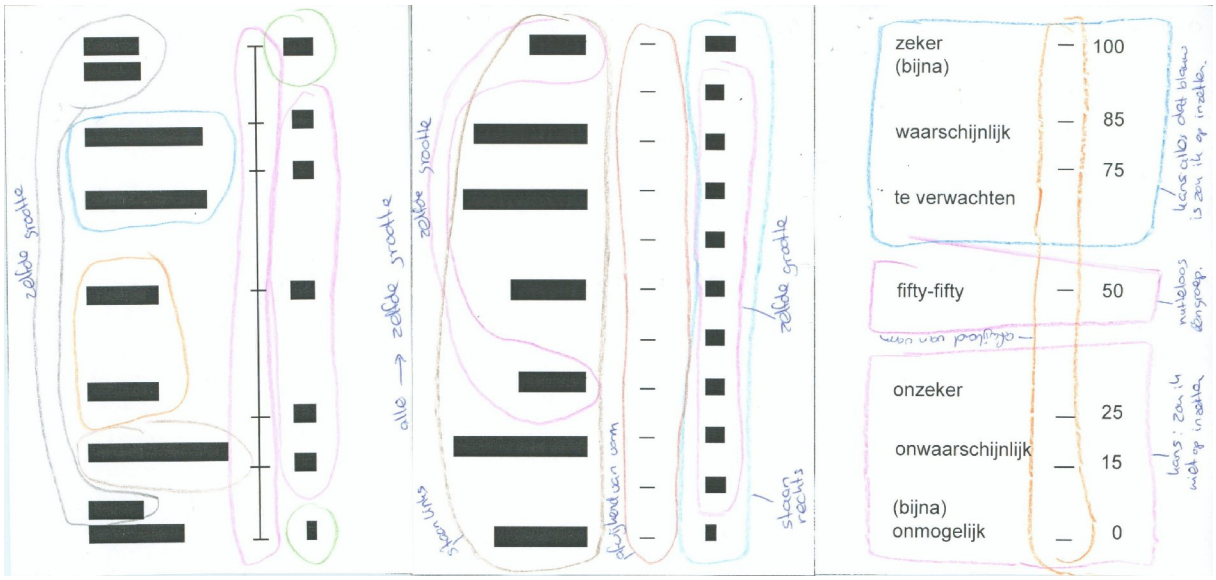
Appendix I: Sheets Experiment 2



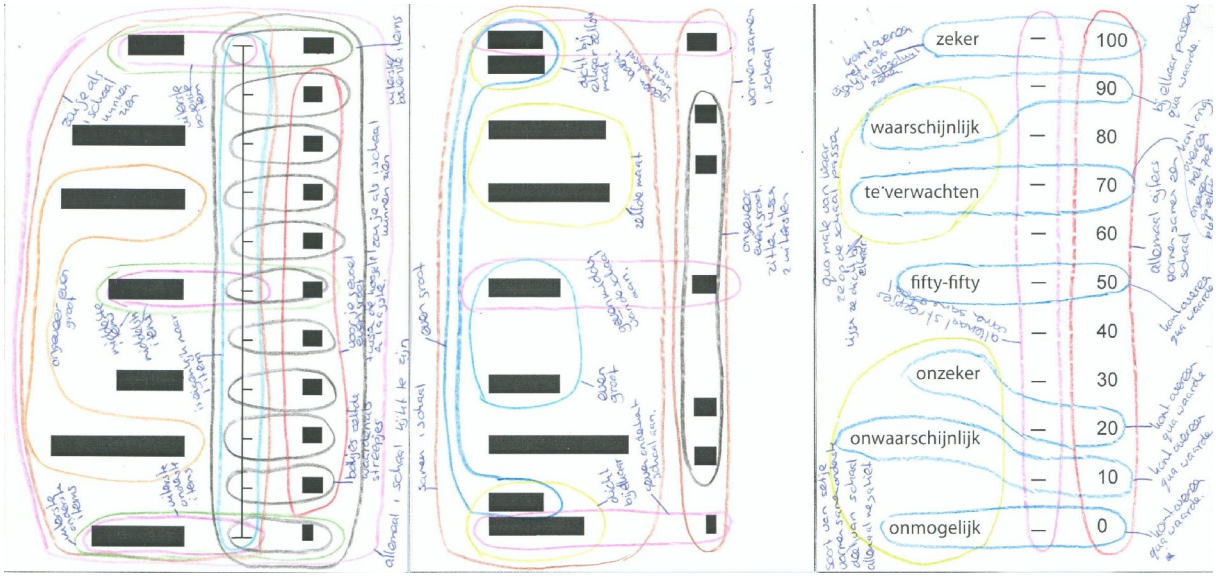
Sheet of Subject 1: 1.1/O2-S3-R3, 1.2/O1-S3-R1 and 1.3/O2-S1-R1



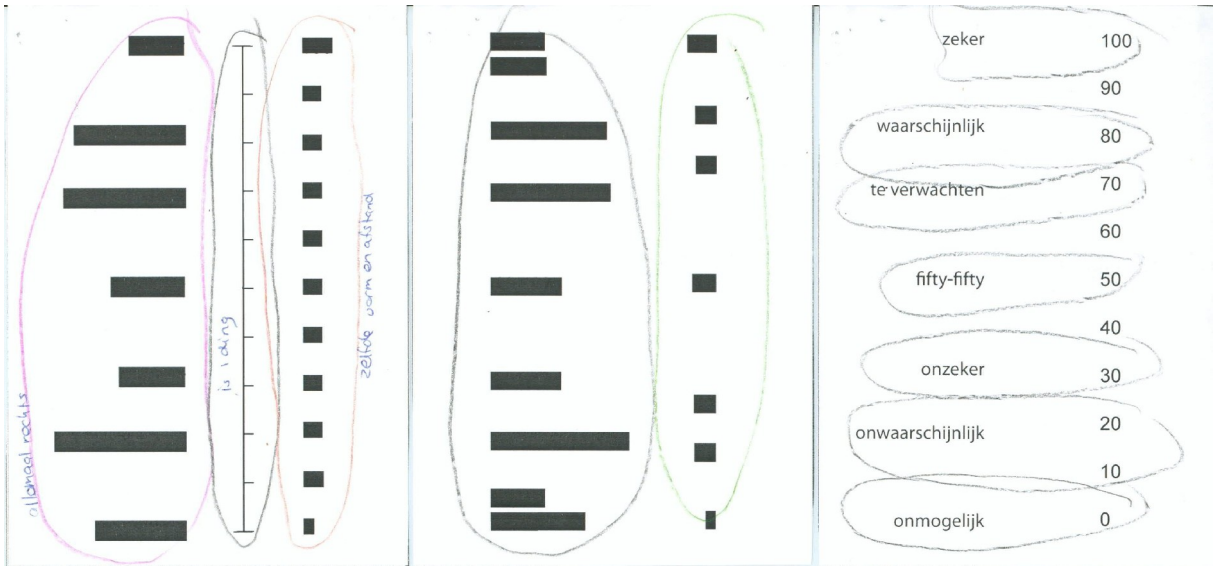
Sheet of Subject 2: 2.1/O1-S3-R1, 2.2/O2-S3-R1 and 2.3/O1-S1-R1



Sheet of Subject 3: 3.1/O1-S3-R1, 3.2/O2-S3-R2 and 3.3/O1-S1-R2



Sheet of Subject 4: 4.1/O2-S3-R1, 4.2/O1-S3-R3 and 4.3/O2-S1-R2

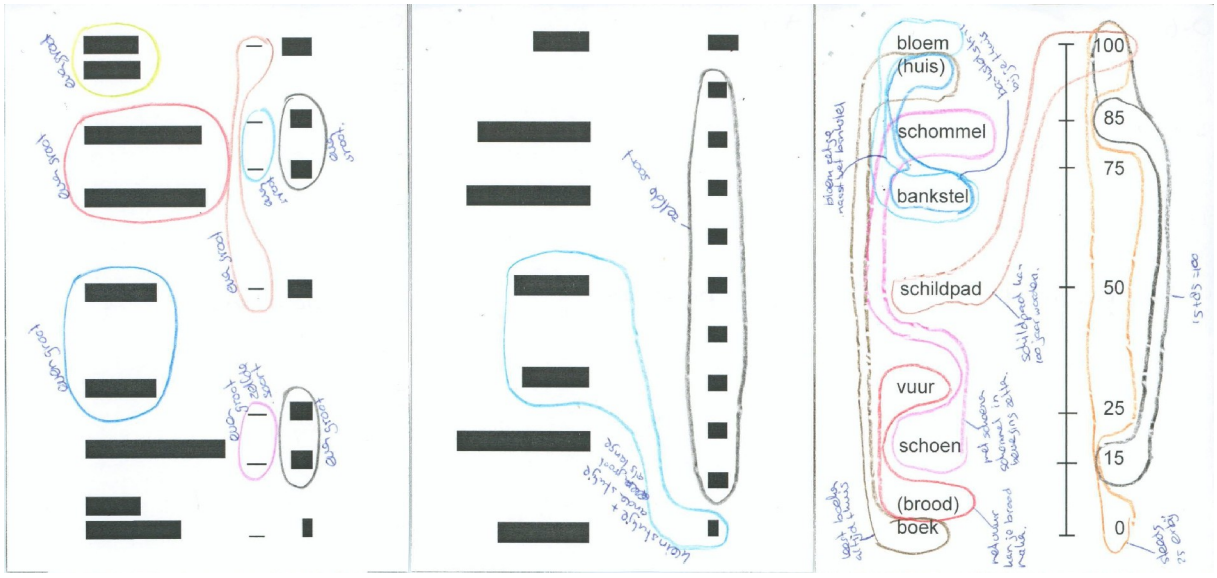


Sheet of Subject 5: 5.1/O2-S3-R1, 5.2/O1-S3-R3 and 5.3/O2-S1-R3

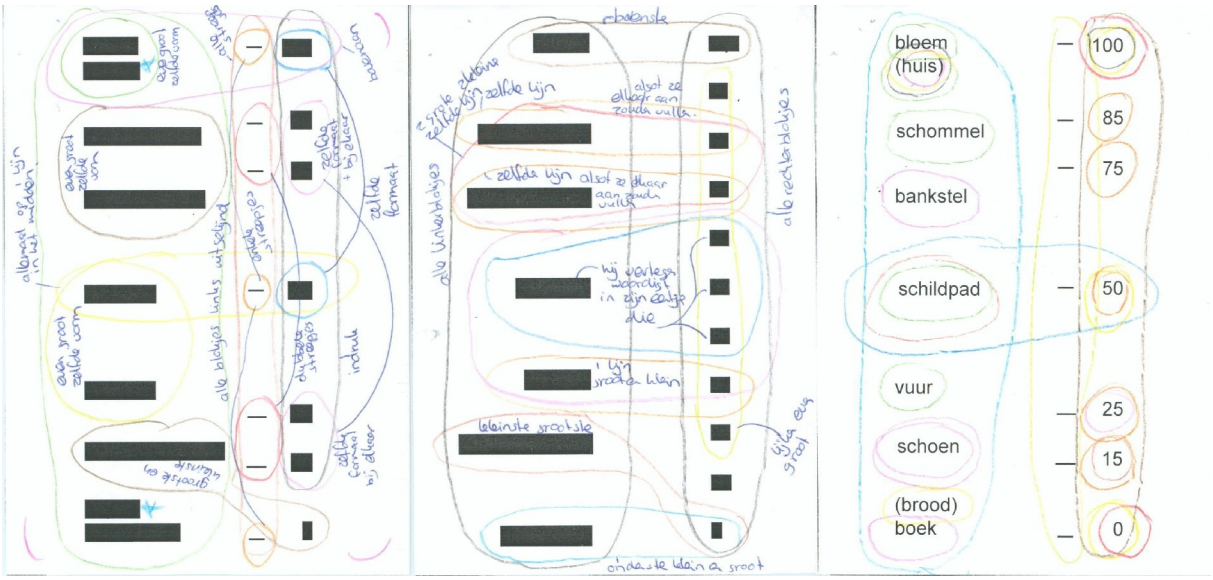
Sheet of Subject 6: 6.1/O2-S3-R2, 6.2/O1-S3-R3 and 6.3/O2-S2-R3

For sheet 6.1/O2-S3-R2 is was decided not to count the clustered objects as a cluster. This subject clustered the objects he believed could not be clustered with an object on the left.

Sheet of Subject 7: 7.1/O1-S3-R2, 7.2/O2-S3-R3 and 7.3/O1-S2-R3



Sheet of Subject 8: 8.1/O1-S3-R2, 8.2/O2-S3-R3 and 8.3/O1-S2-R1



Sheet of Subject 9: 9.1/O1-S3-R2, 9.2/O2-S3-R3 and 9.3/O1-S2-R2