LEARNING IN A MULTIPLE-CUE JUDGMENT TASK:
EVIDENCE FOR SHIFTS FROM RULE BASED PROCESSING TO
SIMILARITY BASED PROCESSING

Joakim Bergqvist

Supervisor: Linnea Karlsson
Department of Integrative Medical Biology
Umeå University
LEARNING IN A MULTIPLE-CUE JUDGMENT TASK:
EVIDENCE FOR SHIFTS FROM RULE BASED PROCESSING TO
SIMILARITY BASED PROCESSING

Joakim Bergqvist

Cue abstraction (additively combining abstracted values) and exemplar memory
(comparing with stored memory via similarity) are important processes in multiple
cue judgments, but previous studies lack insight into how people use these
processes while learning to make judgments. The present study investigates the
learning process in multiple cue judgment tasks, comparing a linear structure with a
non-linear and modeling participant responses with formal models. Concurrent
verbal reporting (think aloud) was used. The hypotheses were a) that initial learning
would follow a "rule bias" via additive integration, b) the representation of the task
would shift to an exemplar memory based one with learning in the non-linear
structure and c) the think aloud protocol would reflect this hypothesized shift. The
multiplicative environment enables better learning of the material, and is best
described by an exemplar memory model while the linear group performs worse
and is equally well described by both models. Model fit in the non-linear group
changes from equal to favoring exemplar memory with training. Hypothesis a was
not supported in the results, both b and c were supported. Furthermore the results
have implications on the question of Rule Bias, and also corroborates previous
studies.

In everyday life we often encounter situations where we need to perform some
sort of judgment, be it how much is reasonable to pay for a certain car or whether
a patient in a hospital should be diagnosed with a certain disease or not. For such
so called multiple-cue judgments it has been shown that (at least) two different
types of processes are involved; exemplar memory and cue abstraction (Erickson &
Kruschke, 1998, Juslin, Olsson & Olsson, 2003). Exemplar memory (Medin, and
Schaffer, 1978; Nosofsky & Johansen, 2000) involves retrieving memory traces of
similar specific instances of a stimulus when making a judgment. For instance, you
might remember seeing a car that looks like the one you are trying to value and
remember the price of that car when you set a value to the present car. On the
other hand, there is cue abstraction where a person uses knowledge of specific
cues, for instance how different parts of a car contribute to the total price
(Einhorn, Kleinmuntz, & Kleinmuntz, 1979).

Several researchers have now begun to shed light on what factors promote the
reliance on these different processes (see e.g. Bröder, Newell, and Platzer, 2010;
Juslin et al., 2003; Juslin, Karlsson, and Olsson, 2008; Karlsson, Juslin, and Olsson,
2008; Von Helversen, Mata, and Olsson, 2010). The main factors behind reliance on
exemplar based memory (EBM) over cue abstraction (CAM) can be attributed to a
multiplicative cue combination (Juslin et al., 2008), a deterministic criterion (Juslin
et al., 2003), or having to retrieve cue information from memory (Bröder et al.,
2010). Bröder et al. (2010) argue that these factors in any combination should
trigger reliance on EBM. However, little is known about the interplay of these processes (CAM and EBM) when learning a judgment task. More specifically, is there evidence for representational shifts between CAM and EBM as learning to make judgments progresses? The purpose of this thesis is to test hypotheses derived from a theoretical framework for multiple-cue judgment called “Sigma” (Juslin et al., 2008), namely that a) irrespective of the certain factors of a task there is an inclination to favor CAM over EBM in the beginning of learning a task, b) if the task does not allow for good performance with CAM (as in a multiplicative task, see below) there will be gradual shifts from CAM to EBM during learning and c) learning to make judgments is a controlled explicit process, and thus it should be possible to capture the shift between CAM and EBM using concurrent think-aloud protocols, where the participant verbalizes how the judgments are made as learning unfolds. In what follows, the framework of Sigma will be spelled out, together with a more specific treatment of the hypotheses.

Sigma as a framework for judgment
Juslin et al. (2008) proposed a framework for multiple cue judgments that utilizes both the concept of the cue abstraction model (CAM) and exemplar based memory (EBM) in a dynamic division-of-labor fashion. The framework is called Sigma (for “summation”). In their article they describe the judgment process as a “controlled cognitive process constrained to serial and additive integration” (p. 263). This “controlled cognitive process” can although it is constrained to additive integration, produce accurate judgments even in tasks where the environment is clearly non-linear or multiplicative (Juslin et al., 2008; Juslin et al., 2003; Karlsson et al., 2008; Olsson, Enkvist, and Juslin 2006). A non-linear or multiplicative environment means that the pieces of information used for a judgment do not all add to the total criterion to be judged in a simple linear fashion.

According to Sigma, CAM uses abstracted values or weights assigned to specific pieces of information (cues), which is appropriate to use in a linear additive environment. EBM, on the other hand is a method of judgment reliant on similarity, which is supposedly used to cope with a non-linear environment. The similarity process compares two items in regards to how many cues differ and combines this with a value of how important that specific cue is to produce the judgment (see below and Eq. 2) (Juslin et al., 2008).

The framework Sigma suggests that it is possible to model the judgment process both when it is fed with abstracted cues (CAM) and when fed with concrete exemplars (EBM). When supplying abstracted cue values, Sigma suggests that the judgment process integrates these sequentially by considering the subjective weight of the cue in relation to the cues previously considered and adjusts the estimated criterion accordingly. When supplying exemplar memory traces, the judgment process according to Sigma compares these traces with stored exemplars with regards to similarity. Multiple exemplar memory traces are compared sequentially and the subjective estimate of the criterion to be judged is calculated by adjusting the estimate in the direction of the retrieved trace. The
magnitude of this adjustment is dependent on the similarity of the trace to the exemplar that is supplied, relative to the other exemplars that are previously attended (Juslin et al., 2008).

Sigma conforms to the notion of a “rule-bias” (Ashby, Alfonso-Reese, Turken, and Waldron, 1998; Juslin et al., 2003; Bröder et al., 2010; Karlsson et al., 2008), in that participants will initially try to infer abstract rules about how the cues and the criterion of the thing to be judged relate, and will only as a back-up resort to using exemplar based reasoning. This implies that when performing a multiple-cue judgment on items based on an additive cue structure, the framework Sigma suggests that CAM should play a major part throughout the judgment process. While when performing a multiple-cue judgment on items with a non-linear structure CAM should dominate at first and progressively change to a reliance on EBM, when the attempts at abstracting cue weights fail.

Furthermore there are a number of predictions that can be made on the basis of Sigma. First of all the items that are presented during training (“old”) and new previously un-encountered items presented during testing (“new”), will have different response patterns depending on which process has been used. A CAM process predicts no significant differences when it comes to response accuracy on new and old items. This due to the fact that the cue weights have been abstracted and can thus be implemented on any item irrespective of whether or not the item has been presented previously or not. When utilizing an EBM process on the other hand the differences between old and new items are predicted to be significantly different. This due to the fact that stored exemplars will result in near perfect accuracy, while un-encountered items are not stored at all and thus have to be judged solely according to their similarity to the stored items. If the participant is encountering a non-linear environment as well, the similarity weights may not be accurate, and might indicate a response that is very wrong. These qualitative patterns can be used to strengthen a modeling result.

Representational shifts in judgment
Juslin et al. (2008) propose that people shift from cue abstraction representations to exemplar based memory representations due to an inability to successfully abstract the linear relations between cues that give satisfactory judgments. Thus, depending on the cue combination rule in effect one should be able to see a clear shift from a CAM representation to an EBM representation (Juslin et al., 2008). For instance, if the task has non-linear cue-criterion relations such as multiplicative relations, that should induce EBM.

That there is a shift from cue abstraction to exemplar memory in a task with a non-linear cue combination rule, as described by Sigma (Juslin et al., 2008) has been tested in a number of studies with convincing results (Juslin et al., 2003; Karlsson et al., 2008; Juslin et al., 2008; Bröder et al., 2010; Von Helversen et al., 2010). This body of research has mainly focused on one training phase followed by one test phase (Juslin et al., 2003; Juslin et al. 2008; Karlsson, Juslin, Olsson, 2007). In these
experiments participants’ performance are measured in the final test and modeled with data from late stages in training. But what happens during the learning phase itself? Is the representational shift evident early or late during learning?

Representational shifts during training have been studied in the related fields of categorization (Johansen & Palmeri, 2002) and subjective probability (Nilsson, Olsson, and Juslin, 2005). Although these studies have found evidence of representational shifts from a cue abstraction model to an exemplar-based model during the actual training phase (Johansen & Palmeri, 2002; Nilsson et al., 2005), these results may not apply to multiple cue judgments. Categorization concerns the ability to perceive an item and choose one category that this item belongs to (Ashby et al., 1998) while multiple cue judgments concerns weighting together multiple pieces of information to judge a criterion, e.g. the toxicity of a bug (Juslin et al., 2003). Thus the processes in use during categorization and multiple cue judgments may in fact be different. The fact that previous studies on multiple cue judgments have shown a number of factors to be successful in inducing an EBM way of solving a multiple cue judgment task (eg. Bröder et al., 2010; Juslin et al., 2003) demands that the potential shift in representations be investigated in a multiple-cue judgment task as well.

The question that becomes evident is thus if there actually are representational shifts from an initial CAM way of solving a task to an EBM way of solving the task, as an effect of a multiplicative environment, when performing a multiple cue judgment task. The hypothesis is that there will indeed be representational shifts when a non-linear cue structure is used. This hypothesis is in line with the previous research presented above as well as the framework Sigma (Juslin et al., 2008). Also, another question presents itself, namely if the modeling data shows the same pattern as data gathered from a verbal protocol performed on a part of the group performing the multiplicative task.

Verbal Protocols
Verbal protocols are effective process-tracing tools, when applied correctly. Ericsson and Simon (1980) proposed verbal reports as a data source, but argued that there are a number of qualitatively different ways of producing introspective verbal reports. One difference can be seen as a temporal aspect. Either a participant produces verbal reports while actively completing the task or the participant supplies a verbal report after a task is completed. These versions of verbal protocols are called concurrent verbal protocols and retrospective protocols respectively (Ericsson & Simon 1980). Kuusela and Paul (2000) compared concurrent and retrospective verbal protocols in order to discern which is better at revealing aspects of human decision making. They found that a concurrent protocol yielded more coded segments than a retrospective protocol did, and thus a concurrent protocol would be more useful if the aim of the study is to examine the process of the decision making.
Concurrent verbal reports can also differ. For instance, there is one version where the participants are asked to explain their choices, constituting Explanatory Verbalization. On the other hand, when participants are only asked to voice their mental speech, the verbal protocol is referred to as a Think-Aloud protocol (Fox, Ericsson, and Best, 2011). Fox et al. conducted a meta-analysis of a large body of studies that used verbal protocols. Their aim was to discern whether a Think Aloud protocol had any effect on the performance of the task at hand. Contrary to the opinion of some researchers (Fox et al. 2011) the result of the meta-analysis showed no significant difference in performance when comparing silent and think-aloud groups. (Fox et al. 2011). Fox et al. did find that there was a significant difference when it comes to time frame of task completion, with longer times for the group performing think-aloud. This is explained as inherent in the verbalization process, verbal speech is slower than mental speech.

On the other hand, Fox et al. found that when comparing results in a setting where explanatory verbalization was used there were significant differences in performance and the explanatory verbalization led to increased performance.

Fox et al. (2011) furthermore emphasize that there are inherent problems with verbal protocols, namely that verbalization is only possible of thoughts that enter our conscious minds. Verbalization of implicit processes should therefore be impossible (Ashby et al., 1998; Fox et al. 2011).

It has been argued that exemplar memory is an implicit process, and should therefore be hard to verbalize (Ashby et al. 1998). However, according to the framework Sigma, the judgment process is a controlled process (Juslin et al., 2003) and there should therefore be indications of explicit controlled judgment not only with cue abstraction but also with exemplar memory. It is therefore expected that a qualitative shift in the representation of the task is to be found in the verbal protocol data. CAM demands that the participant abstracts weights, which should lead to statements where either specific amounts or magnitudes are used in relation to different cues, while EBM emphasizes on similarity. Thus expected EBM expressions should refer to similarity, or recognition of a previously stored exemplar (Juslin et al., 2003; Juslin et al., 2008).

The aim of the present study when it comes to verbal protocol data is therefore to discern whether it is possible to express EBM-type expressions in a think-aloud task and whether these possible expressions resonate with the formal modeling results. Previous research indicates that expressing exemplar-type expressions should be very hard, and that verbal reports should therefore conform to a rule based type in line with CAM (Ashby et al., 1998). The hypothesis is on the contrary that modeling data will support a shift from CAM to EBM in a multiplicative environment but not in an additive environment where both models will be able to describe the performance, while verbalization data in a multiplicative environment should follow exemplar memory type expressions.
To summarize, in this study it will be investigated if there are representational shifts from CAM to EBM during training in a multiple-cue judgment task with a non-linear cue structure, as well as if a concurrent think aloud protocol follows the same pattern of shifting. It is hypothesized that a) There will be an initial bias to rely on rule abstraction, namely a CAM way of solving the problem, b) there will be a shift from a reliance on CAM to a reliance on EBM as an effect of training in the multiplicative environment and c) the think aloud data will follow the same pattern of shifting from CAM to EBM.

The Experiment
The present paper describes an experiment designed to explore the quantitative model fits of participants during the course of training in a multiple-cue judgment task through the use of intermediate testing during training. This design is a modification of the design used by Juslin et al. (2003). The task in question was to learn to judge the toxicity of a fictitious bug, the Death Bug. This bug varied on four weighted binary cues that together with an intercept value produced the criterion to be judged. The experiment was implemented as a between-group design, where one group encountered a linear additive combination of the cues and the other a nonlinear multiplicative combination of the cues. Both groups encountered deterministic criteria. To further investigate the hypothesized representational shift a part of the multiplicative group performed a concurrent verbal report. This report was conducted as think aloud, where the participant verbalizes internal speech. Furthermore Sigma predicts specific patterns of judgment for items that are learned and new items for both CAM and EBM. A person using CAM when judging will have no systematic differences between old and new items. While a person utilizing EBM will have significant differences between old and new items (Juslin et al., 2008).

To formally model the participants’ responses during both intermediate tests and final testing, a cue abstraction model and an exemplar-based model were used. This was done to enable a comparison of model fits as an effect of training.

Thus the experiment explores the effects of training in a multiple cue judgment task, with both an additive and multiplicative combination rule.

Method

Participants
Fifty participants took part in the study, aged between 18 and 36 ($M=23.8, SD=3.3$). 20 of these were women and 30 were male. Due to technical errors two participants were excluded, giving a final number of 48 participants, 19 women and 29 male. Participants recruited were mainly undergraduate students at Umeå University. All participants were informed that the test was voluntary and could be aborted at any time. All participants also signed an informed consent paper.
For their participation, all participants received a payment of at least 75 SEK, and an additional bonus depending on their performance. The participants who performed the think aloud protocol were given an initial payment of 100 SEK, and an additional bonus. The bonus received was dependent on which task environment the participant encountered, an additive or a multiplicative task environment. This was due to the fact that the multiplicative task environment was expected to require more time to complete, based on previous research results stating that an additive environment is easier to learn (Juslin et al., 2003). If the participant encountered an additive task environment, the maximum bonus was 50 SEK, while if the encountered task environment was multiplicative the bonus was doubled. The bonus was calculated using the performance scores of the participant. Root mean square error (RMSE) scores (see eq. 3) on the first intermediate test, the last of the performed intermediate tests and the final test were used to calculate the bonus. The calculated RMSE was categorized into one of four groupings. See table 1 for exact rewards and groupings depending on performance. The reward from the final test was double that of the intermediate tests.

<table>
<thead>
<tr>
<th>Environment</th>
<th>RMSE≥4</th>
<th>4&gt;RMSE≥2</th>
<th>2&gt;RMSE≥1</th>
<th>1&gt;RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Add</td>
<td>6.13</td>
<td>8.17</td>
<td>9.80</td>
<td>12.25</td>
</tr>
<tr>
<td>Mult</td>
<td>12.25</td>
<td>16.33</td>
<td>19.60</td>
<td>24.50</td>
</tr>
</tbody>
</table>

*Reward measured in Swedish Kronor, SEK

**Design and Material**

A between-group design was used, in which half of the participants encountered a linear additive task environment and the other half encountered a non-linear multiplicative task environment (see Juslin et al., 2008, for two similar tasks). In addition, eight of the participants encountering the multiplicative environment were instructed to perform concurrent verbal reporting (think-aloud) while executing the task.

The participants’ task was to judge the toxicity of a fictitious bug, the Death Bug. The Death Bugs’ toxicity varied depending on four binary cues, c₁-₄, producing a cue structure with 16 combinations, see Table 2. In the additive environment the toxicity (i.e. the criterion) was determined by a linear function,

\[ C = 10 + 20 \cdot c_1 + 15 \cdot c_2 + 10 \cdot c_3 + 5 \cdot c_4 \]  

(1)

The criterion, C, for the multiplicative environment was determined by a non-linear equation,

\[ C = 9 + \left( \frac{20c_1 + 15c_2 + 10c_3 + 5c_4}{12,7} \right) \]  

(2)
These cue-combination rules were chosen in order to produce a deterministic criterion of equal range for both task environments during testing, while the range varied during training. The full range of the criterion for both the additive and the multiplicative environment can be found in Table 2. The physical cues of the task were balanced in that participants were randomly divided into eight groups and each group was assigned a physical cue set.

Table 2. Exemplars with cue structure and criterion values.

<table>
<thead>
<tr>
<th>Exemplar No.</th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$C_3$</th>
<th>$C_4$</th>
<th>Additive</th>
<th>Multiplicative</th>
<th>Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>60</td>
<td>60</td>
<td>Final test</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>55</td>
<td>44</td>
<td>Training</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>50</td>
<td>32</td>
<td>Intermediate</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>45</td>
<td>32 (25)</td>
<td>Training</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>45</td>
<td>25</td>
<td>Final Test</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>40</td>
<td>20</td>
<td>Training</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>35</td>
<td>16</td>
<td>Intermediate</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>14</td>
<td>Training</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>40</td>
<td>20</td>
<td>Training</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>35</td>
<td>16</td>
<td>Intermediate</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>30</td>
<td>16 (14)</td>
<td>Training</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>12</td>
<td>Final Test</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>25</td>
<td>12</td>
<td>Training</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>20</td>
<td>11</td>
<td>Intermediate</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>15</td>
<td>10</td>
<td>Training</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>10</td>
<td>Final Test</td>
</tr>
</tbody>
</table>

Training: Items presented during training and testing. Intermediate: Items presented during intermediate tests and final test. Final Test: Items presented only at the final test.

Procedure

Before the training started the participant was presented with some fictive background information about the bug in question. The participants were instructed in how the experiment would be conducted, specifically the layout of training blocks and intermediate tests. The participants were also instructed that the length of the experiment depended on their performance. Participants were instructed that the toxicity of the bug was measured in percent and that they should guess if they did not know how poisonous the bug was. They were also told how the reward was calculated.

In each trial the participant’s task was to judge the toxicity of the subspecies presented in Table 2. The subspecies varied on four binary cues; Yellow or grey.

---

1 Due to an implementation error, two items (exemplar no. 4 and 11) received a criteria that deviated from the expected value, see Table 2 for factual values (expected value within parenthesis).
head, red or blue back, long or short legs and small or big eyes. Weights assigned to each visual cue were ordered in eight different sets, of which one was randomly selected for each participant. Presentation order was individually randomized for each participant.

The participant encountered a picture presented on a computer screen together with the question: “How poisonous is this Death Bug?”. The participants answered by typing in a numerical value using a regular keyboard connected to the computer. After each answer the participant was presented with the correct answer in a feedback slide, showing both the participant’s own answer and the correct answer along with the image of the bug. During the intermediate tests and the final test no feedback slide was shown.

The experiment was divided into a number of training blocks and intermediate tests as well as a final test. The duration of the experiment varied as an effect of the number of training blocks and test phases presented to the participant, dependent on the learning rate. All participants completed at least three training blocks, consisting of a total of seven exposures to each of the 8 training items. Each participant also encountered two sub-tests as well as the final test. Beyond that the participants could encounter seven additional training blocks and additional two sub-test phases. Each of the seven additional training blocks, except for one, consisted of two exposures of each training item, while the one consisted of one exposure of each training item. This generated a total of at least 56 training trials for each participant and a total possible number of trials of 160 (see figure 1).

![Test Layout Diagram]

*Figure 1.* Description of test layout, with number of exposures of each item within parenthesis.

To discern whether a participant had learned sufficiently or not the RMSE value of each training phase was calculated. If this value was below 1.5 the participant was
presented with the final test (see Von Helversen et al., 2010, for a similar learning criterion). A low RMSE score on the training blocks indicate a high level of proficiency in the task. To guarantee that the mandatory three training blocks were completed by each participant the RMSE calculation was conducted on training blocks 4 and onwards (see figure 1). The reason for such a learning criterion was to ensure that the participants had similar levels of skill when they were presented with the final test.

The intermediate test phases consisted of two exposures of each item, while the final test consisted of three exposures of each item. During the intermediate test phases all training items as well as four new items, marked as intermediate in table 2, were presented. The final test contained all of the items previously encountered, training and intermediate, as well as the items marked as Final test in table 2.

Eight of the participants encountering the multiplicative environment performed a Think Aloud protocol. This protocol was recorded digitally. The test supervisor was present in the room while the participants performed the think aloud, in order to both prepare the participants and also to prompt the participants to continue with the think aloud, should they cease to think aloud. The supervisor was seated behind the participant in an unobtrusive way.

Before the experiment started the participants were instructed in how to perform the think aloud as well as trained in performing the protocol. The participants were instructed to say whatever came up in their heads, without shortening or summarizing their thoughts. Explanations of why the participant thought in a certain way were also discouraged, unless the explanation was not due to the think aloud but a part of the original thought. See appendix A for the warm-up exercises and instructions used for the think aloud protocol.

*Dependent Measures*
Throughout the experiment a number of measures were gathered pertaining to the performance of participants. For every participant root mean square error (RMSE) values were calculated for every trial in every training block and test block. These RMSE values are calculated through the use of the participants’ estimates of the criterion, specifically by the use of equation (3)

\[ RMSE = \sqrt{\frac{\sum_{i=1}^{n} (Correct_i - Response_i)^2}{N}} \]  

where Correct corresponds to the correct criterion value and Response is the judgment supplied by the participant and N the number of trials. Consequently a low RMSE score indicates a good response that deviates very little from the correct answer.

The participants’ estimates of the criterion were also used in the modeling. Participants’ replies during testing were fed into the modeling equations (cue
abstraction model and exemplar model, see Appendix B) in order to calculate a root mean square deviation (RMSD) value output from the model. This value is calculated in a similar way as equation (3) but the input is the judgment supplied by the modeling equation and the participant response instead of a correct criterion and a judgment. This results in a measure of how well the model fits the participants' data.

Furthermore the number of training blocks completed before achieving the target accuracy of an RMSE of <1.5 was recorded.

*Expected Statements During Think Aloud*

In order to investigate the hypothesized shift in representation of the participants a verbal protocol analysis was conducted on eight of the participants performing the task in the multiplicative task environment. A set of expected statements were produced in advance in order to operationalize the process and enable quantification of the participants’ statements.

The expected statements were grouped into three different categories, with a number of typical sentences each. The categories were “Cue Abstraction, With Numbers” (CAM-NUM), “Cue Abstraction, With Quantities” (CAM-QUA) and “Exemplar Memory” (EBM). A very strict classification was enforced, to ensure that the actual observation pertains to the particular process. This was hypothesized to exclude a large amount of statements that fall into a “grey zone”, being classifiable as cue abstraction and exemplar memory type expressions simultaneously, if a less strict classification was enforced.

The CAM-NUM category captures the process of abstracting exact values of a cue, corresponding to a strict cue abstraction process. The CAM process is thought to be an explicit process (Juslin et al., 2003), and thus statements of the sort captured by the CAM-NUM category are clearly signs of a cue abstraction way of solving the task at hand.

The category of CAM-QUA type expressions capture the essence of cue abstraction in that it requires magnitudes in relation to a specific cue, thus ensuring that the abstraction of a cue weight has been achieved. The process itself is explicit, but the relative weights may not be uttered. The participant still has an internally represented linear weight assigned to the magnitude, even if it is not uttered specifically (Juslin et al., 2003).

Expressions categorized as EBM rely on the concept of similarity. If the participant expresses recognition of a previously seen item, that indicates that previous instances of said item are being compared to the present item. In order to ensure that only very clear Exemplar memory processes were captured by the protocol, a very strict way of classifying the EBM statements was utilized. This had the effect that only expressly clear or exact recognition of an item was considered as EBM expressions, together with sudden insight into the actual weight of the item. This
kind of sudden insight is hypothesized to correspond to the process of trying to abstract and weigh together cues, but suddenly realizing that the similarity of the item is in fact great enough for exact recognition. Also, an explicit statement of not looking at the separate cues, but at the whole was also considered as an EBM statement, since such a statement clearly indicates that the participant is cognizant of the fact that he or she is not performing cue abstraction but rather trying to memorize or learn the wholes instead.

See table 3 for a more detailed list of requirements for classification for each category, together with expected sample sentences.

While listening to the recorded think aloud the scorer gave one point to each separate category for each time that condition was fulfilled, with a maximum of one point per category per item.

The score was then averaged individually for each participant over two presentations of each item (16 trials). This was due to technical limitations, which only permitted discriminating training items and corresponding verbal expressions as being between two test phases. This information coupled with the number of training blocks completed by the participant enabled a comparison of the average number of expressions per category per analysis phase. An analysis phase was defined as the training blocks between two test phases (see fig. 3).

**Results**

*Performance During Training*

During training participants performed judgments on the stimuli in a number of training blocks. RMSE values were calculated for all training blocks using equation (3). Average performance for each block is shown in Table 4.

Although these results do not in themselves concern the hypotheses, they are relevant because of other aspects. Learning speed differences and differences between the participants performing think aloud and those not performing it could have impacts on consequent results and discussions.

In order to determine if there was a difference in learning speed between the two task environments, the two groups were compared with regards to number of blocks completed before being presented with the final test. A one-way ANOVA with number of blocks completed before achieving the training criteria as dependent variable and task environment as between-subjects factor revealed that there was a trend towards that the additive environment was slower in reaching the training criteria, but the test did not reach significance \[ F(1,46) = 3.4; \text{MSE} = \]

Table 3. *Statement classes with descriptions and expected example sentences.*
<table>
<thead>
<tr>
<th>Statement Type</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAM-NUM</td>
<td>A cue coupled with a specific value</td>
<td>&quot;The eyes, they give 10 in toxicity.&quot;</td>
</tr>
<tr>
<td></td>
<td>Arithmetic expressions</td>
<td>&quot;Five plus ten, but not the eyes, so minus 15. It should be 20.&quot;</td>
</tr>
<tr>
<td></td>
<td>The difference between a specific cue’s binary state (0/1) expressed with a number</td>
<td>&quot;So, big eyes instead of small add 10 to the toxicity.&quot;</td>
</tr>
<tr>
<td>CAM-QUA</td>
<td>A specific cue coupled with a quantity and direction of change</td>
<td>&quot;The eyes, they give a lot of toxicity.&quot;</td>
</tr>
<tr>
<td></td>
<td>Indicating direction of the criterion based on a single cue's binary state</td>
<td>&quot;So, if the eyes are big instead of small raises the toxicity.&quot;</td>
</tr>
<tr>
<td>EBM</td>
<td>Clear recognition</td>
<td>&quot;I've seen that one before. It had around 50.&quot;</td>
</tr>
<tr>
<td></td>
<td>Exact recognition</td>
<td>&quot;That's the most toxic one.&quot;</td>
</tr>
<tr>
<td></td>
<td>Sudden insight or realization</td>
<td>&quot;... but it's THAT one! That one has 45.&quot;</td>
</tr>
<tr>
<td></td>
<td>Explicitly stating that the main focus is the overall item instead of cues</td>
<td>&quot;I'm just looking at them as wholes now.&quot;</td>
</tr>
</tbody>
</table>
Table 4. Judgment performance during training, intermediate tests and final test as measured by Root Mean Square Error (RMSE) between criterion and judgment.

<table>
<thead>
<tr>
<th>Task</th>
<th>Additive</th>
<th>Multiplicative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase</td>
<td>Mean</td>
<td>Std. Dev</td>
</tr>
<tr>
<td>Block 1</td>
<td>16.32</td>
<td>3.76</td>
</tr>
<tr>
<td>Block 2</td>
<td>10.39</td>
<td>2.96</td>
</tr>
<tr>
<td>Block 3</td>
<td>8.02</td>
<td>3.07</td>
</tr>
<tr>
<td>Block 4</td>
<td>7.25</td>
<td>3.80</td>
</tr>
<tr>
<td>Block 5</td>
<td>6.80</td>
<td>3.53</td>
</tr>
<tr>
<td>Block 6</td>
<td>6.53</td>
<td>2.93</td>
</tr>
<tr>
<td>Block 7</td>
<td>5.93</td>
<td>3.94</td>
</tr>
<tr>
<td>Block 8</td>
<td>8.77</td>
<td>13.87</td>
</tr>
<tr>
<td>Block 9</td>
<td>5.32</td>
<td>4.95</td>
</tr>
<tr>
<td>Block 10</td>
<td>5.40</td>
<td>4.56</td>
</tr>
<tr>
<td>Test 1</td>
<td>12.95</td>
<td>2.30</td>
</tr>
<tr>
<td>Test 2</td>
<td>11.38</td>
<td>3.08</td>
</tr>
<tr>
<td>Test 3</td>
<td>10.52</td>
<td>3.45</td>
</tr>
<tr>
<td>Test 4</td>
<td>9.20</td>
<td>3.75</td>
</tr>
<tr>
<td>Final Test</td>
<td>11.96</td>
<td>4.77</td>
</tr>
</tbody>
</table>

14.1; \( p = .07 \). The trend in the results disappear when excluding the group performing Think Aloud [\( F(1,38) = 0.5; \text{MSE} = 0.15; p = .83 \)].

Following this result, a comparison of how many participants there were in each group that reached the training criterion was conducted. In the additive task environment group 11 of 24 (46%) of the participants achieved the criteria, while 17 of 24 (66.6%) in the multiplicative task environment group reached the criteria. However, a Chi\(^2\) test, environment by number of participants reaching the criteria, only approached significance [\( \chi^2(1) = 3.1; p = .08 \)]. When excluding the think aloud group the results were even more similar with 9 of 16 (56%) [\( \chi^2(1) = 0.42; p = .52 \)] in the non-think aloud multiplicative task environment group reaching the criteria. In the think aloud group every participant reached the training criteria.

Contrary to this result the multiplicative task environment group performed better on the last training block performed. A one-way ANOVA with performance on the last block, measured in RMSE, as dependent variable and task environment as the independent variable shows that the multiplicative task environment group performed significantly better [\( F(1,46) = 4.5; \text{MSE} = 54.8; p = .039 \)]. This significance does not remain when the think aloud group is excluded from the comparison [\( F(1,38) = 1.9; \text{MSE} = 27.7; p = .17 \)].
In sum, the results on performance during training demonstrate that while the additive task environment and the multiplicative task environment groups learn equally fast, the multiplicative task environment group performs the task significantly better at the end of training. Also the results indicate that the think aloud group contributes markedly to the effects shown in the analyses above.

**Performance on Intermediate tests**

Next was to investigate how the participants performed during the intermediate tests, using the RMSE values for the intermediate tests. Again, these results do not in themselves pertain to the hypotheses, but are relevant nonetheless. Further investigation of possible differences between the think aloud and non-think aloud groups are of interest.

In a repeated measures ANOVA on participants who performed all four intermediate tests, with task environment as between-subject factor and intermediate test as within-subject factor, there was a main effect of both group \( [F(1,36) = 4.9; \text{MSE} = 126.1; p = .033] \) and intermediate test \( [F(2.11,36) = 18.35; \text{MSE} = 210.88; p = .000] \) but no interaction effect \( [F(2.11,36) = 0.98; \text{MSE} = 11.23; p = .385] \), with the multiplicative group performing better than the additive group. Note that the calculations for intermediate test and the interaction effects with that factor violate the sphericity assumption, and thus the degrees of freedom have been corrected in lieu with Greenhouse-Geisser.

This indicates that performance increased for every subtest and that the multiplicative task environment group was significantly better than the additive task environment group. These effects remain when excluding the think aloud group, meaning the main effect of group \( [F(1,33) = 4.6; \text{MSE} = 119.8; p = .039] \), main effect of intermediate test \( [F(2.11,33) = 14.2; \text{MSE} = 173.16; p = .000] \) and no interaction effect \( [F(2.11,33) = 0.67; \text{MSE} = 8.16; p = .522] \). Again, these calculations have been likewise been corrected with Greenhouse-Geisser.

**Performance on Final Test**

Performance on the final test was measured as the RMSE between the correct criterion and the participant’s estimation. This was done in order to investigate the patterns of the responses to see if they were in line with what Sigma predicts. When comparing RMSE values for so called *old items* (items encountered during training) and *new items* (items presented only at the final test; see Table 2) as within-subjects factor and task environment as between-subjects factor, with a repeated measures ANOVA there was a main effect of item type \( [F(1,46) = 180.28; \text{MSE} = 3522.16; p = .000] \) and an interaction effect of item type and task environment \( [F(1,46) = 4.05; \text{MSE} = 79.03; p = .05] \), but no main effect of environment \( [F(1,46) = 0.001; \text{MSE} = 0.021; p = .981] \). The main effect of item type remains when removing the Think Aloud group from the comparison \( [F(1,46) = 120.41; \text{MSE} = 2594.44; p = .000] \) but the interaction effect is no longer significant \( [F(1,46) = 1.56; \text{MSE} = 33.69; p = .219] \).
This indicates that the performance on old items is better than new items in both groups. While the interaction effect of task environment and item type is significant \( (p = .05) \), a follow up one-way ANOVA showed no significant differences in performance between the groups when it comes to old items \( [F(1,46) = 2.48; \text{MSE} = 40.82; p = .122] \), although it approached significance, or for new items \( [F(1,46) = .92; \text{MSE} = 38.23; p = .343] \). This shows that both groups perform equally well on both types of items, which is unpredicted by Sigma. Sigma predicts large differences on performance of new items, where participants performing the test in the additive task environment are predicted to perform much better (Juslin et al., 2008).

*Cognitive Modeling of the Judgment Processes*

In order to model the participants’ judgment processes both an exemplar based memory model and a cue abstraction model were employed (see Appendix B for a description of the mathematical formulations of said models). In order to control for over-fitting a leave-one-out cross-validation procedure was used (Stone, 1974; Von Helversen et al., 2010). The modeling, when using cross-validation, works by using participants’ responses from the test sequences. The models are fitted to all but one item, in order to estimate the free parameters of the models. These estimated parameters are then used in order to predict the participant’s response on the left out item. For the intermediate tests 11 items are used to predict one item, while correspondingly 15 items are used in the final test (see table 2). This process of estimation and prediction is then repeated for every item in the test phase. In order to calculate the goodness of fit, the predicted responses are compared to the participants’ responses (as averaged across the total exposures of each item in each test). The resulting discrepancy is measured in root mean square deviation (RMSD) between the actual response and the predicted response.

For the cue abstraction model five parameters were estimated, an intercept and four cue weights (see equation 1). These parameters were estimated using a Simplex algorithm as implemented in Matlab. This algorithm finds the parameters that produce the lowest output of a function, in this case the RMSD value between the factual and predicted responses. The starting values for the Simplex algorithm were produced by randomly assigning values within the range of the weights.

For the exemplar memory model four free parameters were estimated in the same manner, corresponding to the cue structure similarity between probe and exemplar. These cue structure similarities then impact the similarity weight of the equation, \( S_n \) in appendix B2. The starting values for the Simplex algorithm in the exemplar memory model were assigned by randomly producing a value in the interval \([0,1]\).

The modeling of the cue abstraction model was conducted five times per participant, while the exemplar memory modeling was conducted 100 times, per
participant. This difference was due to the fact that the optimal weights for the cue abstraction model could be estimated regardless of the value of the starting parameters, while the multiplicative nature of the exemplar model made the search for optimal parameters much dependent on the exact starting values.

**Participant classification**

Participants were classified into either CAM or EBM depending on the RMSD of their modeling results. This was done for each individual test that had been modeled. The classification of the present study’s participants was done where the difference between the models, in regards to RMSD, was at least 0.2 or greater. This differentiation was chosen because this value represented the smallest standard error of either of the tests and models, thus providing a strict classification criterion. See table 5 for an overview of amounts of participants classified into each model, per test and table 6 for model fit values. If the difference between the modeling results were less than 0.2, the participant was deemed unclassifiable for the test in question.

Table 5. *Number of participants classified as CAM, EBM or Unclassifiable (NONE) per modeled test*

<table>
<thead>
<tr>
<th>Environment</th>
<th>Intermediate Test 1</th>
<th>Intermediate Test 2</th>
<th>Final Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CAM</td>
<td>EBM</td>
<td>NONE</td>
</tr>
<tr>
<td>Additive</td>
<td>5</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>Multiplicative</td>
<td>4</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>34</td>
<td>5</td>
</tr>
<tr>
<td>Think Aloud*</td>
<td>1</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

*Think Aloud also included in Multiplicative.

Table 6. *Model fit (RMSD values, mean and standard deviation), averaged over environment.*

<table>
<thead>
<tr>
<th>Environment</th>
<th>Test 1</th>
<th>Test 2</th>
<th>Final Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CAM</td>
<td>EBM</td>
<td>CAM</td>
</tr>
<tr>
<td>Additive</td>
<td>Mean</td>
<td>Std. D.</td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td>8.49</td>
<td>3.58</td>
<td>7.40</td>
</tr>
<tr>
<td>Multiplicative</td>
<td>7.68</td>
<td>3.21</td>
<td>6.50</td>
</tr>
<tr>
<td>Mean</td>
<td>8.09</td>
<td>3.39</td>
<td>6.95</td>
</tr>
</tbody>
</table>

Due to time constrains only the first two intermediate tests could be modeled, as well as the final test.
Results of formal modeling

In the multiplicative task environment group, 3 non-think aloud participants were classified as CAM and 12 as EBM for the first test, 2 as CAM and 14 as EBM for the second test and 1 as CAM and 15 as EBM for the final test. The think aloud part of the multiplicative task environment group had 1 classified as CAM and 7 as EBM for the first test, 0 as CAM and 8 as EBM for the second test and 0 as CAM and 8 as EBM for the final test. A comparison of these two groups, a Chi Square test, showed no significant differences in the amount of participants classified as either CAM or EBM, between the two groups (think aloud/no think aloud) in the first intermediate test \( \chi^2(1) = 0.105; p = .622 \), the second intermediate test \( \chi^2(1) = 1.091; p = .356 \) or the final test \( \chi^2(1) = 0.522; p = .667 \). When comparing the additive and the multiplicative task environment groups using a Chi Square test, there were no significant differences in the amount of participants classified into CAM and EBM (see table 5 for exact numbers) depending on task environment for the first intermediate test \( \chi^2(1) = 0.206; p = .721 \) or the second intermediate test \( \chi^2(1) = 1.113; p = .396 \). The results of the Chi Square for the final test approached significance \( \chi^2(1) = 4.181; p = .062 \). Which indicates that there was almost a significant difference in classifications between the two environments. Note that a large group of the participants performing the task in the additive task environment were unclassifiable (see table 5). Also see figure 2 for an overview of participant classifications for the three tests that were modeled, averaged over task environment and table 6 for numerical averages.

In order to determine that the modeling of the multiplicative task environment group did not differ significantly between the think aloud group and the non-think aloud group, a repeated measures ANOVA test with modeling RMSD results for CAM and EBM, and test phase as within-subjects factor and think aloud as between-subjects factor, on the participants in the multiplicative group, there was no main effect of group \( [F(1,22) = 0.44; \text{MSE} = 5.23; p = .512] \). This enables a comparison of the entire multiplicative group with the additive group.

Enabled by the result of the previous analysis, a comparison was made between the additive and the multiplicative task environment groups in regards to model fit. This comparison was done in order to investigate if there are systematic differences in how well the models fit the participants depending on the task environment and if this fit changes over time. Thus the additive and the multiplicative task environment groups were compared with a repeated measures ANOVA with test and model fits as within-subjects factor and task environment as between-subjects factor. There were main effects of Test \( [F(2,46) = 33.28; \text{MSE} = 420.38; p = .000] \), of Model \( [F(1,46) = 79.84; \text{MSE} = 363.38; p = .000] \) and task environment \( [F(1,46) = 4.95; \text{MSE} = 84.39; p = .031] \). There were also interaction effects for test and task environment \( [F(2,46) = 7.87; \text{MSE} = 99.44; p = .001] \), model and task environment \( [F(1,46) = 37.29; \text{MSE} = 169.74; p = .000] \), test and model \( [F(1.28,46) = 13.01; \text{MSE} = 74.66; p = .000] \), as well as test, model and task environment \( [F(1.28,46) = 14.54; \text{MSE} = 83.45; p = .000] \). The interaction effect of
model and task environment is not apparent in intermediate test 1 \[F(1,46) = 0.022; \text{MSE} = 0.05; p = .883\], but in intermediate test 2 and the final test, the interaction effect of model fit and task environment is significant \[F(1,46) = 27.53; \text{MSE} = 220.22; p = .000\] \[F(1,46) = 35.08; \text{MSE} = 56.27; p = .000\]. Also note that the interaction of test and model showed significance on Mauchly’s test of sphericity, and thus are corrected according to Greenhouse-Geisser.

Furthermore an identical repeated measures ANOVA was conducted, with one change, namely that the groups were tested individually. This in order to investigate the pattern of model fit shift separately for both groups. The part of the repeated measures ANOVA that is of interest is thus the interaction effect of model and test. Both tests violated the Sphericity assumption for the interaction of model and test, and are thus corrected according to Greenhouse-Geisser. For the additive task environment group the interaction of model and test showed no significance \[F(1.42,23) = 1.85; \text{MSE} = 3.12; p = .181\]. The multiplicative task environment group on the other hand did show significance for the interaction effect of model and test \[F(1.22,23) = 16.11; \text{MSE} = 161.92; p = .000\]. Which indicates improved fit for the EBM model as training progresses.

These analyses indicate that the EBM model and the CAM model fit equally well initially, only to change over time. The multiplicative task environment group is better described by an EBM model, while the additive task environment group is equally well described by both, as shown in figure 2. Another result to note is the main effect of task environment, which indicates a significant difference in the model fits of the two groups, with EBM fitting the multiplicative task environment group better than the additive task environment group and CAM fitting the additive task environment group significantly better.

![Figure 2](image_url)

*Figure 2. Mean RMSD values per test for the two task environments, error bars represent 1 Standard Error.*

**Think Aloud Results**

Think aloud statements were rated as either being included in one of the three categories or not. This was done for every participant that was performing think aloud. Due to time constrains this was only performed by one rater.
As stated previously due to technical constraints it was not possible to discern exactly which of the 10 training block a participant was performing when a specific statement occurred. Despite this it was possible to divide the entire session into five different analysis phases, a phase defined as being before a sub test. The number of item presentations between each intermediate test differed according to a set pattern (see figure 1). After each intermediate test one repetition of each item was added to the training phase. In training phase 1 each item was presented 2 times, constituting 16 items. In training phase 2 each item was presented 3 times, constituting 24 items, etc. (see figure 3).

Due to this, averaging the number of statements per analysis phase was done in line with analysis phase 1, meaning the number of statements in each category was averaged over two repetitions of each item. This was done to enable comparing mean number of statements per phase per participant. This was furthermore clouded by the fact that every participant did not perform every training block in each phase, due to achieving the training criteria. This was accounted for in the averaging of statements. Mean number of statements per analysis phase is presented in Table 7.

![Figure 3](image-url) Layout of the analysis phases with corresponding number of training items per analysis phase.

As shown in figure 4, the number of statements that are classified as EBM increase in proportion towards the later analysis phases. To be taken into account is that the number of participants towards the later phases decreases since more participants achieve the training criteria.

Also, see Appendix C for an excerpt of the verbal protocol.

**Discussion**

The goal of the present thesis was to investigate the processes underlying multiple cue judgments during the training of a multiple cue judgment task. The main question was whether there is evidence for a shift from a cue abstraction way of solving the task to an exemplar memory based model, as an effect of training, in a multiplicative task environment. This was investigated by utilizing a formal modeling based on the framework called Sigma (Juslin et al., 2008), as well as a concurrent verbalization of inner speech, Think Aloud (Eriksson and Simon, 1980).
Table 7. Number of statements during think aloud, means and SD for each analysis phase averaged over 16 items.

<table>
<thead>
<tr>
<th>Type</th>
<th>1</th>
<th>Mean</th>
<th>Std. Dev</th>
<th>2</th>
<th>Mean</th>
<th>Std. Dev</th>
<th>3</th>
<th>Mean</th>
<th>Std. Dev</th>
<th>4</th>
<th>Mean</th>
<th>Std. Dev</th>
<th>5</th>
<th>Mean</th>
<th>Std. Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAM-NUM</td>
<td>0.3</td>
<td>0.7</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>2.1</td>
<td>0.4</td>
<td>0.8</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAM-QUA</td>
<td>0.4</td>
<td>0.5</td>
<td>1.3</td>
<td>1.3</td>
<td>0.1</td>
<td>1.4</td>
<td>0.9</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EBM</td>
<td>1.5</td>
<td>1.8</td>
<td>3.8</td>
<td>3.8</td>
<td>1.6</td>
<td>3.0</td>
<td>7.6</td>
<td>2.6</td>
<td>10.7</td>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4. Distribution of statements for each analysis phase. N=number of participants performing the phase in question.

The hypotheses were that irrespective of task environment, the initial process would be abstract rule generation (CAM). Modeling data would support a shift from a CAM process to an EBM process in the multiplicative task environment group, but there would be no such shift in the additive task environment group and they would be equally well described by both models throughout the task. Finally the think aloud data would follow the modeling data, since the exemplar based memory process is considered a controlled process of judgment (Juslin et al., 2003).

The first hypothesis of the present study concerned whether or not participants would start off following a “rule bias” (Ashby et al., 1998). The hypothesis was that participants would show an initial rule bias, irrespective of which environment the participant encountered. The results show a different pattern.

The modeling of the participants’ judgment processes shows that there is a greater reliance on EBM than on CAM during the initial phases of training. After only 16 items, two repetitions of each item, the EBM model fit more participants where previous research states that there should be greater reliance on CAM (Juslin et al, 2003; Juslin et al., 2008). This pattern is evident in both the additive and the multiplicative task environment. This is in direct opposition of the “rule bias” (Ashby et al., 1998; Juslin et al., 2003) that is expected from previous studies. Which calls into question whether there is such a bias in this type of task.
Could it be that initial attempts to abstract cues fail and thus a greater weight is put on EBM, or is it that these two processes develop in parallel initially? Looking at the additive task environment group the model fit of both models improves with training, but they are equally good throughout the test. This would indicate that both models are suited for this type of task. It could also be that these types of processes work in unison. During a judgment, it might be that exemplars of extremes are stored and then these criteria are adjusted according to the cue structure and cue weights.

Regarding the second hypothesis, that the modeling data would support a shift from CAM to EBM in the multiplicative non-linear task environment, there is evidence of precisely that.

The data indicates that the models’ ability to describe a participant’s responses changes with training. The models yield significantly different results, with EBM describing the participants better, in the multiplicative task environment group. This shows that already after 40 items (the first two training blocks) the EBM model performs significantly better than CAM in the multiplicative task environment group. One could argue that this does not enable the use of the phrase “shifts from rule based processing to similarity based processing”. That is incorrect though, since the modeling of the initial intermediate test shows a similar RMSD for both models. These RMSD values are actually quite high, both of them, which indicates that neither of the models fit the multiplicative task environment group that well initially. This changes over time with training to indicate a significantly better fit for the EBM model. The fit of the EBM model in the final test (M=2.8) is actually a rather good fit, while the fit of the CAM model is very bad (M=6.1). This indicates that the participants rely more on an EBM process and less on a CAM process as the test progresses, thus indicating a shift from a rule based process to a similarity based process.

It would have been interesting to see the modeling results of the other test phases, but as stated previously, due to time constraints only the first two tests could be modeled, as well as the final test. The trend of better performance of the EBM model continues into the final test, which indicates increased learning during the subsequent training phases.

The fit of the model is significantly better with each measurement and the difference in model fit between the EBM and CAM model is significant from the start. In the multiplicative task environment group only 1 person is better explained by a CAM model in the final test, which indicates clear reliance on the EBM model in this sort of environment. These findings clearly support previous findings in this field (Juslin et al., 2003; Juslin et al., 2008; Karlsson et al., 2008) and highlights the fact that this pattern of representational shifting occurs very early on in the training.
When scrutinizing the participant classifications, it should be noted that there is a general trend towards a greater reliance on an EBM model than on CAM. This is expected in the multiplicative task environment, where CAM is notoriously hard to achieve to a satisfactory level (Juslin et al., 2003). However previous studies have reported better fit with CAM than with EBM for the additive task environment group, and thus greater reliance on such a model (Juslin et al., 2003; Juslin et al., 2008). Furthermore the response patterns on old and new items also follow the pattern of an EBM way of solving the task, as predicted by Sigma, which gives additional credence to the modeling data. This can be partially explained by the deterministic criterion, which is known to promote exemplar memory (Juslin et al., 2003).

However, Bröder et al, (2010) argue that there are a number of factors promoting EBM, but that they do not in themselves facilitate a shift in representation. For instance a factor such as being unable to satisfactorily abstract cue values (Juslin et al., 2008) together with a deterministic criterion could initiate a shift to EBM even in an additive environment. Could this inability of the participants in the additive task environment group to satisfactorily abstract the cue values be a reason for the decreased performance, as well as the reason for most participants to rely on an EBM process? If so, the natural question is what the reasons behind this inability are.

Regarding the think aloud data and thus the third hypothesis, the concurrent verbal report supports the modeling data, in that EBM typical statements constitute an increasingly large proportion of the expressions as training progresses. This follows the hypothesis that exemplar memory, as a controlled process should be verbalizable. This finding in particular is interesting, in that this result goes against the previously stated theory that exemplar memory is an implicit process (Ashby et al., 1998) and should thus be very hard, if not impossible, to verbalize. The results of this experiment give clear indications that this is not the case. Although these results clearly indicate that verbalizing EBM is possible, it would be interesting to see whether a verbalization of the additive group would result in the same pattern or if it would follow a more CAM oriented way of expressing the process.

This furthermore strengthens the assumption of Sigma, namely that Exemplar Based Memory is a controlled process, which one relies on when the ability to perform an abstraction fails or does not perform satisfactorily (Juslin et a., 2008).

Besides the hypotheses there are a number of findings worth noting. Firstly there is the fact that some of the statistical significances disappear when removing the think aloud group. This indicates that these participants perform the task better in some ways (discussed below). This effect does not occur when scrutinizing the modeling data, which indicates that the modeling of the participants in the multiplicative task environment group is independent of their actual performance on the tests, furthermore strengthening the idea of greater reliance on EBM when
performing a task with a multiplicative combination rule in a multiple-cue judgment task.

Also the performance data during intermediate tests show that there is clear evidence that the multiplicative task environment group performs the task better than the additive task environment group. The data from the intermediate tests show significant differences in RMSE values depending on task environment, but no interaction effect of task environment, indicating that the multiplicative task environment group and the additive task environment group learn the task equally fast, but that the multiplicative task environment group outperforms the additive task environment group. This goes against previous studies (eg. Juslin et al. 2008) that support faster learning in an additive task environment.

The increased performance of the multiplicative task environment group is also unpredicted, since non-linear tasks are generally described as more difficult to learn and master (Juslin et al., 2003; Olsson et al., 2006)). The increased performance on the tests could be ascribed to the fact that intermediate tests were used, where the multiplicative task environment group could have stored exemplars of the test items, and thus increasing their performance on later repetitions of said exemplars. It should be noted that this difference remains when excluding the think aloud group.

The think aloud group’s increased performance is entirely unpredicted. According to the meta-analysis performed by Fox et al. (2011) concurrent verbalization of inner speech should not alter performance on the task performed. This is not the case in the current study, where performance increased significantly in the think aloud group. The difference in performance could be ascribed to Logan’s Instance Theory (Logan, 1988), which assumes that each encounter with an item encodes it in memory and retrieves previously stored encounters, all by the same act of attention. Could it be that the think aloud task altered the attention of the participants in a way that improved the storing or possibly the retrieval of the item? Logan states that both storing and retrieval of an item “is an obligatory, unavoidable consequence of attention” (p. 493 Logan, 1988), and that this storing and retrieval may be better or worse, depending on the task and the situation. Thus it could be that the task of thinking aloud increases the participants’ attention on the item and therefore the quality of storing and/or retrieval increases in a way that improves the participant’s performance on the task at hand. No previous experiments with a multiple-cue judgment task utilizing concurrent verbalization were found prior to conducting this experiment, and thus the question remains if this difference in performance is due to the nature of the task or if it is an artifact dependent on some factor of the procedure of the current experiment.

The outcome of this thesis indicates the need of more research in this field. Specifically there is need of a study that utilizes think aloud in an additive task environment as well, something that was excluded from the present study due to time constraints. The verbal reports from such a study could shed light on the
process of judgment where cue abstraction is actually expected, hence indicating if people do abstract cues in the way Sigma suggests. Statements from such a study would possibly shed light on the question of a possible relational combination of the models. Furthermore the results of test performance in the additive task environment group would indicate if this type of task gains from performing verbalization of inner speech or if a gain is only evident in a multiplicative environment.

Furthermore a similar study utilizing a probabilistic criterion would give insight regarding the issue of a possible lack of rule bias and the obvious reliance on EBM in the additive task environment group perceived in the present study. A probabilistic criterion would predict better learning in an additive environment when using CAM and impaired learning when trying to use EBM (Juslin et al., 2003) and thus one would possibly see the expected rule bias. If not, then the present study's results are further strengthened.

Also, studies utilizing concurrent verbal reports in a more structured way would furthermore shed light on the effects of verbal reports on multiple-cue judgment tasks. Further studies of this kind would give an indication if the results of Fox et al., (2011) are valid or if the findings in the present study are valid.

Conclusions

In sum, the results of the conducted experiment indicate that there is a) no indications of a "rule bias" in the current experiment, rather the data support an initial reliance on an EBM process, b) that there is indeed a shift from a cue abstraction model of judgment towards an exemplar based memory process as an effect of training in a multiplicative environment and c) concurrent verbal reporting of inner speech follows the pattern that exemplar-based modeling data defines.

References


Appendix A. Warm-up procedure of think aloud.

Used with permission from Dr. Edward Cokely, Michigan Technological University.

**Verbal Protocol – Concurrent Think Aloud Instructions**

*Always sit behind or otherwise out of view of the participant-
Be sure recording equipment is unobtrusive*

**Read these instructions:**

I will start by familiarizing you with the procedure for “thinking out-loud.” We are interested in knowing your thoughts as you come up with the answers to the problems in this experiment. In order to do this, I am going to ask you to think out-loud as you work on some practice questions. What I mean by “think out-loud” is that I want you to say your thoughts out loud from the moment you finish hearing a practice question until you say the final answer. **Don’t summarize what you say,** in fact, it doesn’t even need to make sense. Just think out loud as if you are alone and speaking to yourself. Let’s’ start with an easy question:

**What is the first letter after A**

B is correct. When I asked you “what is the first letter after A,” you might have thought of B, but felt unsure of whether some thought occurred in between. In that case, you’d only say B (like you did). Only report thoughts that you have. Don’t try to report unclear thoughts or thoughts you think you should have had or might have had, just say your thoughts out loud as if you were alone, as much as you comfortably can. And remember- never summarize your thoughts. Let’s try another example- Please think out loud:

“What is the sixth letter after B?”

Good. The letter “H” rarely immediately comes to mind after hearing the question. People commonly need to go through several steps to find the answer. However, when someone summarizes their thinking, **which we don’t want,** they might say “I am finding the letter H by counting through the alphabet.” Instead, when people solve this problem by thinking out loud, they usually say a sequence of individual letters, such as A B, C, D, E, F, G, before they answer H. Remember, you never need to summarize your thoughts, just say exactly what you’re thinking “out-loud.”

Ok, let’s try another practice question. Are you ready? Think aloud while you generate the answer. How many dot are in this array

(Experiment- present 27 dot grid)

(Experimenter, give feedback on participant’s answer, if needed remind them to
(Experimenter- present 2 dot grid)

(Experimenter- ask the participant to describe the difference between thinking out loud and summarizing thoughts)

If need be, the experimenter may administer the following questions for added practice:
- “How many months begin with the letter J?”
- “What is the fifth letter before M?”
- “name 5 animals in a zoo”
- What is 14*9
- “What is 13*4
- “Name 3 different team sports”
Appendix B1. Formal modeling formula, CAM

The linear additive model was used to model the cue abstraction data,

\[ \hat{c} = a + b_1 \times C_1 + b_2 \times C_2 + b_3 \times C_3 + b_4 \times C_4 \]

where the intercept, \(a\), and the linear weights \(b_n\) are estimated by the Simplex algorithm as implemented in Matlab. \(C_n\) refers to the cue state of the relevant cue (1 or 0). When an optimal abstraction of the linear weights, \(b_m\), and the intercept \(a\), has been achieved the linear additive model is a perfect match to equation (1), and thus produces perfect estimates of the criterion.

Appendix B2. Formal modeling formula, EBM

The formal modeling formula used in the experiment the exemplar model presented by Juslin et al. (2003) in relation to the model Sigma

\[ \hat{c} = \frac{\sum N S_n \times c_n}{\sum N S_n} \]

where \(\hat{c}\) refers to the estimated value of the criterion to be judged. \(N\) refers to the number of exemplars. \(S_n\) is the similarity between the probe, the stimulus at hand, and the exemplar stored in memory. \(c_n\) refers to criterion of the exemplar \(n\). To calculate the similarity between the probe and the exemplar stored in memory the original context model’s similarity rule was used (Medin & Schaffer, 1978).

\[ S_n = \prod_{i=1}^{l} d_i \]

This equation produces a similarity based on \(d_n\) which takes the value of 1 if the cues on the dimension \(i\) are the same for both the probe and the exemplar. If they differ \(d_i\) takes on the value of \(s_i\), the impact values of deviating similarity, in the range of \([0, 1]\). These values were estimated by the Simplex algorithm in Matlab.
Appendix C. Excerpt from think aloud data.

Original statements in Swedish, translation provided by the author.

- “Här har vi i alla fall den allra giftigaste, den hade 44.”

Translation: “Here at least we have the most toxic one, it had 44.”

This sentence was classified as being an EBM type expression, since it clearly fulfills the requirement of exact recognition.

- “Och här har vi en med giftiga ögon och giftiga vingar och det är den första som var med och det är 20 i giftighetsgrad på den.”

Translation: “And here we have one with toxic eyes and toxic wings, which is the first one that was in the test and the toxicity is 20 for that one.”

This sentence was also classified as an EBM type expression, since it also fulfills the requirement of exact recognition.

- “Och här kommer högre, 25 plus massa annat. Det blir kanske... ja, vi sätter 35 på den.”

Translation: “And here comes a higher one, 25 plus a lot of other stuff. That might be... well, let’s put 35 on it.”

This sentence was classified as a CAM-NUM, since it fulfills the requirement of an arithmetic expression.

- “Ja, farlig variant, snälla ben, ganska farliga ögon tror jag, då blir det 32 om jag ska testa den teorin.”

Translation: “Right, a dangerous one, harmless legs, pretty dangerous eyes I think, then it is 32 if I’m going to test that theory.”

This sentence was classified as a CAM-QUA sentence, since the eye cue is related to a quantity and a direction of toxicity.