university of groningen / faculty of science and engineering

What Can We Learn from Absent Cues?

Master Project Thesis

Sanne Poelstra, s2901560, s.poelstra.1@student.rug.nl
Supervisors: Dr. J. van Rij-Tange & Dr. J. S. Nixon

Abstract: Error Driven Learning (EDL) is a theory of learning that states that we learn by using stimuli (cues) to predict certain outcomes. EDL is often represented by a simple neural network with cues as input, outcomes as output and weights as predictions. Theories of learning on how to update those weights exist and they differ in their implementation. Rescorla and Wagner (1972, RW) state that absent cues (cues that have been seen before, but are not seen now) should not lead to an update of the weights, while Van Hamme and Wasserman (1994, VHW) propose that absent cues *should* update weights. In this thesis, we modelled the experiment from Van Hamme and Wasserman (1994) with both the RW and VHW algorithms. Although these algorithms make different predictions in certain circumstances, the original experiment does not seem to tease these two predictions apart. For that reason, we conducted three experiments. We found support in the direction of the Rescorla-Wagner model, thus indicating that learning might not occur in the absence of cues. We will discuss the results in terms of task effects on implicit learning versus explicit inference and how this aspect could be addressed in future research.

1 Introduction

"[Y]ou begin learning in the womb and go right on learning until the moment you pass on" (Gelb & Buzan, 1996, p. 16). When we first come into this world all the input that we receive is new. There are shapes that we have to make sense of, and sounds that might or might not be important. How do we learn all these new relationships and how do these relationships change over time?

According to Error Driven Learning (EDL) theory (Rescorla & Wagner, 1972; Widrow & Hoff, 1960), learning is a process of minimising uncertainty about upcoming states in the world. We do this by using incoming sensory information (*cues*) to predict upcoming events (*outcomes*).

EDL has been shown to work in multiple domains of learning, such as first language acquisition (e.g.: Hsu et al., 2011; St. Clair et al., 2009), second language learning (Ellis, 2006) and other parts of linguistics (e.g.: Arnold et al., 2017; Nixon & Tomaschek, 2020). As well as fields such as age research (Ramscar et al., 2017), developmental psychology (e.g.: Ramscar, Dye, Gustafson, & Klein, 2013; Ramscar et al., 2007). This indicates that EDL is not constrained to explaining only one part of learning.

EDL is often represented in a simple neural network such as the one seen in Figure 1.1. This network has an input and output layer and is fully connected, without hidden layers. The input layer represents the cues and the output layer represents the outcomes. This is an asymmetric network, as the cues predict outcomes, but not vice versa.

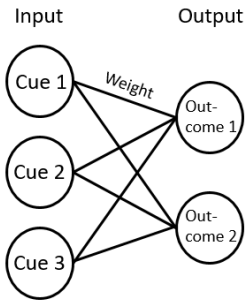

Figure 1.1: Simple EDL network with an input (cues) and output (outcomes) layer, fully connected.

A connection from one cue to one outcome has a certain weight, this weight indicates how predictive a cue is of an outcome. When the summed connection weights of all present cues to an outcome (called the activation) are close to zero, this represents high uncertainty for that outcome. If several outcomes have equal activation, that could mean high uncertainty if the outcomes are in competition with each other. As learning is a process of reducing uncertainty, updating the weights is a very important feature of this simple neural network. We can model learning by updating the weight according to what cues and outcomes we come across. But how do we update those weights?

This thesis begins by discussing two different theories on how to update the weights in this EDL network. First we discuss a theory where the weights of cues that are absent (not seen) are not updated, after that we will discuss a theory where the weights of cues that are absent are updated. We will also discuss some general characteristics of EDL and compare those to other forms of learning. We will then go on to discuss an experiment performed to test whether weights of absent cues should be updated, based on which we ask the question: Do people learn from absent cues?

1.0.1 Rescorla and Wagner; absent cues are not updated

One of the main theories on how to update the weights is given by Rescorla and Wagner (1972). They state that updating the weights depends on whether or not the cue is present, and that the weights should *only* be updated when a cue is present.

How to update those cues, according to Rescorla and Wagner, is shown in Equation 1.1. ΔV_{ij}^t is the change in weights between cue i and outcome j at a certain time and η is the learning rate (typically set to 0.01). In the original paper η is represented by α and β, where the first is the learning rate for the cue and the second the learning rate for the outcome. However, the form we present here is more similar to the Delta formula (Widrow & Hoff, 1960), which we chose to limit the number of parameters to fine-tune (free parameters) when modelling this process. Act_j^t is the activation of outcome j, which is the sum of the connection weights for the present cues, it lies between 0 and 1.

$$\Delta V_{ij}^t = \begin{cases} 0 & \text{, cue } i \text{ absent,} \\ \eta(1 - act_j^t) & \text{, cue } i \text{ and outcome } j \text{ are present,} \\ \eta(0 - act_j^t) & \text{, cue } i \text{ present but outcome } j \text{ absent} \end{cases} \tag{1.1}$$

We will first look at the last two lines of the equation. The second line states that if a cue and outcome are both present, then there is an increase in weight. This means that the connection between that cue and outcome (the degree to which the cue predicts the outcome) is strengthened. In the last line we see that if a cue is present but the outcome is not, the weight between the two decreases and thus the connection weakens. The exact amount of adjustment to the weight depends on how expected the outcome is, based on all present cues (the activation).

As mentioned before, in EDL cues predict outcomes, therefore Rescorla and Wagner state that if a cue is absent, it is not informative. We cannot predict an outcome if there is no cue present, as there is nothing to base the prediction on. This can be seen in the first line of the equation, which states that if a cue is absent, there is no update to the weights.

1.0.2 Van Hamme and Wasserman; absent cues are updated

While Rescorla and Wagner argue that absent cues should not cause an update to the weights, several researchers did not agree with this notion. For example, Markman (1989) stated that we often actively encode certain features as "missing". Tassoni (1995) also stated that an absence of a stimulus could be information about the correlation between that stimulus (or cue) and outcome as well. One of the main (often cited) theories that agrees with this stance of missing cues being informative, is that of Van Hamme and Wasserman (1994).

Van Hamme and Wasserman propose that accounting for absent cues is something that is necessary in describing EDL. While they do agree with Rescorla and Wagner on what happens when a cue is present, they propose an adjustment to the formula for the absent cues. Their formalisation can be seen in Equation 1.2 and it is based on the work of Markman (1989) mentioned earlier.

In this equation, ΔV_{ij}^t still represents the change in weights between cue i and outcome j at a certain time. Just as Rescorla and Wagner, Van Hamme and Wasserman use an α and β as the learning rates for the cues and outcomes respectively. Here we simplified that again into one learning rate of η. There are now a negative and positive version of η, which is because Van Hamme and Wasserman actively encode a missing cue as *absent*. They do this by multiplying an present cue by 1, and an absent cue by -1. In our use of 0.01 for the learning rate η, this means that we now either have -0.01 $(-\eta)$ or 0.01 (η). This change leads to the formulas on the first two lines having an opposite effect when compared to the formulas on the last two lines.

$$\Delta V_{ij}^t = \begin{cases} -\eta(1 - act_j^t) & \text{, cue } i \text{ absent but outcome } j \text{ present,} \\ -\eta(0 - act_j^t) & \text{, cue } i \text{ and outcome } j \text{ absent,} \\ \eta(1 - act_j^t) & \text{, cue } i \text{ and outcome } j \text{ are present,} \\ \eta(0 - act_j^t) & \text{, cue } i \text{ present but outcome } j \text{ absent} \end{cases} \tag{1.2}$$

The first line in the formula states, if a cue is absent but an outcome is present, the weight decreases (the opposite effect of the third line), and the connection between the two is thus weakened. The second line states that if a cue and outcome are both absent, then the weight increases, so the connection between the two is strengthened (opposite effect of the fourth line). The third and forth line are the same as in the Rescorla and

Wagner formula, so an increase in weight when both are present, but a decrease when a cue is present but an outcome is not.

This thesis investigates if people actually learn from absent cues. Which of the two models discussed best captures the learning trajectory?

Van Hamme and Wasserman (1994) explored part of these questions in their own paper, where they performed an experiment to test their addition to the formula of Rescorla and Wagner. We will discuss this experiment in detail here, but we will first explain more about the general properties of EDL.

1.0.3 General properties of EDL

It is important to establish other main characteristics of EDL and how exactly EDL differs from other forms of learning. One of the main other theories of learning is statistical learning. While this term is used differently depending on the context, we define it as follows: statistical learning models propose that people keep track of the statistical distribution of cues and that learning is based on the actual distribution of the inputs. This in contrast to EDL, which states that the main part of learning occurs when there is a prediction error. While statistical learning is thus more based on the input (cues) itself, EDL is more based on how informative the cues are about the outcome.

The two key phenomena that distinguish these two theories of learning are cue competition (blocking) and unlearning (Nixon, 2020).

First we will talk about blocking. Kamin (1967) showed that the learning of a new cue can be 'blocked' if there is already a previously learned cue that predicts a given outcome strongly enough. If an animal hears a bell and then sees food, there is a connection between the bell and the food. If the animal is then later also introduced to a flashing light in combination with the food, this new cue (the flashing light) is blocked, as the bell is already a strong predictor of the food. Blocking cannot be explained by just a statistical learning account, as the behaviour can not just be explained by looking at the probability of food appearing after either a bell or a light. It is a process where all available cues *compete* to predict the relevant outcome. In the neural network this is represented by the fact that the weights are always updated with respect to the whole system, such that all present cues influence each other's weight adjustment. As we can see in Equation 1.1 for example, we subtract the activation of the outcome from either 1 or 0. The activation is the sum of the connection weights for the present cue. Thus if this sum is very large (because the bell is already very strongly connected), then the adjustment to the weight of the flashing light will be very small.

The second difference between EDL and statistical learning is unlearning. While statistical learning mainly focuses on learning from associations, EDL also focuses on unlearning. In EDL learning is not just about gaining more knowledge, but also about reducing uncertainty about how well these pieces of knowledge predict other pieces of knowledge. Unlearning is about learning to ignore the unreliable cues. As Rescorla (1988) stated, learning depends on how well a certain cue predicts a following outcome, and not just on the amount of times that the cue and outcome occur together. This is reflected in the last lines of Equations 1.1 and 1.2. Here a cue is present, but an outcome is not. Thus the model adjusts itself such that this cue is now less predictive of the outcome. This would ideally result in a model learning what *not* to expect.

One last aspect of EDL to discuss, is that it could be an implicit process. Ramscar, Dye, and Klein (2013) looked at how adults and children learn in the context of EDL. They found that children showed different results in a word learning task than adults, which they attributed to adults applying reasoning strategies, whereas children only seemed to use implicit learning. The reason for this difference between age groups is that the prefrontal cortex, the part of the brain responsible for processes such as logic or reasoning (explicit inference), is not yet fully developed in children. The development of the prefrontal cortex might be responsible for the difference between children and adults, as it allows adults to use explicit task strategies, while children cannot use those yet. Children's results looked more similar to what EDL predicts would happen, while the adults diverted from that prediction.

Implicit learning is best done without interference from logic and reasoning. Therefore explicit inference might interfere with or hinder EDL, if EDL is indeed an implicit process.

1.0.4 Van Hamme and Wasserman's 1994 experiment

Now that we have a better understanding of the workings of EDL, we will discuss the experiment that Van Hamme and Wasserman performed (1994) to test their addition to the Rescorla Wagner model and to see if there is evidence that people learn from absent cues or not.

In the experiment participants had to determine if certain foods would result in an allergic reaction or not. Each participant saw three different sheets, with each sheet consisting of three different foods

(A, B and X). This could, for example, be peanuts, shrimp and yogurt respectively. On this sheet for every trial they had to fill in ratings for all three foods, indicating how likely they thought it was that these foods would lead to an allergic reaction. They rated this causality on a scale from 0 (definitely not) to 8 (definitely). Each trial (or day as it was framed to the participant) presented to the participant a slide with two out of the three foods displayed, together with whether or not there was an allergic reaction that day. In any given trial, food X was always seen and food A and B were seen half the time. This means that in each trial, a participant saw either food A and X or food B and X on the screen, together with an outcome. So in a trial where A and X were shown, the rating for B was of importance as that was the absent cue. In a B and X trial, the rating for cue A was important as here that was the absent cue.

Van Hamme and Wasserman found that if a certain cue was present, the ratings increased over trials when there was an outcome present, and decreased when an outcome was absent. This is in line with the predictions of Rescorla and Wagner as well, and it is what the last two lines in both Equation 1.1 and 1.2 state.

For cues that were not present, if there was an outcome, then the ratings decreased over trials. If there was no cue and no outcome, then the ratings increased. This is in line with the addition to the Rescorla-Wagner formula that Van Hamme and Wasserman suggest.

1.0.5 The current thesis

Van Hamme and Wasserman did not run simulations of their predictions. Therefore they did not have any formal predictions about the difference in performance between their model and that or Rescorla and Wagner. One of the aims of the current study is to simulate the experiment and generate formal predictions on how both models would behave. Modelling can show the behaviour of the cues over the experiment and see if Van Hamme and Wasserman did test what they set out to test with their experiment, or if they might have found an effect of something different.

While using the allergist paradigm is something that is common in EDL research (see Houwer and Beckers (2002) for a review of different studies using this paradigm), it could be the case that people will already have preconceived notions about certain foods influencing the results. Peanuts and wheat for example are foods known for causing allergies (NHS, 2019a) and were shown in the experiment. While they were not in the same food group (A, B or X), Van Hamme and Wasserman never reported on the individual foods and their pre-existing connection to an allergic reaction or the lack thereof. M. Le Pelley et al. (2013) state in their paper that multiple researchers found that participants learn faster about cues they have seen predict something before, than cues that were not predictive, indicating that previous knowledge will indeed matter in the learning process. For this reason in our experiment we would like to avoid using the food paradigm and use new cues and outcomes that should not have a previous relationship to each other, thus investigating whether Van Hamme and Wasserman's theory also applies to learning in other contexts.

This thesis aims to find out which of these two mechanisms best describes Error Driven Learning. Do people learn in the absence of cues?

We will do so by firstly making a computational simulation of the experiment that Van Hamme and Wasserman performed in their 1994 paper. To model both theories we will use a simple implementation of a neural network as described in the beginning of this introduction. Others have also successfully simulated EDL learning with this type of simple neural network, see Hoppe et al. (2020) for an explanation and review.

Based on the results of this simulation we will then perform three different experiments. Our aim was to conduct four experiments, with the last experiment's goal being to investigate implicit learning with a forced choice paradigm. However due to time limitations we were unable to run all four experiments. We will discuss this latter experiment in the discussion.

Our first two experiments changed the stimuli compared to the original Van Hamme and Wasserman experiment, however as we did not fully replicate their findings with these two experiments, our third experiment was a direct replication of the original experiment (including the food stimuli). As we do still think the food paradigm might influence the results, we will introduce extra measures in this experiment to investigate the influence of the connection between food and allergic reactions. Each experiment will be discussed in their own section, after which there will be a global discussion at the end of the thesis.

2 Computational Modelling

To verify the advantages of making trial-by-trial predictions through modelling, we developed simulations of both the Rescorla-Wagner model and the Van Hamme-Wasserman model. We investigate in which ways these two models differ in predictions, and which one of them more accurately predicts the behaviour seen in the original Van Hamme and Wasserman experiment.

Before presenting our simulations, we first explain the experiment of Van Hamme and Wasserman in detail. We will then end this section with the results of these simulations of the experiment and compare them to the results of the original experiment as well.

2.1 The 1994 experiment

The participants in the experiment were told to imagine that they were an allergist (someone who is trained to manage and treat allergies), who was trying to determine the cause of an allergic reaction shortly after their patient ate something. There were three different types of food a participant was presented with during a block. These were encoded as foods A, B and X and were always shown as compound cues. Compound cues are cues that are shown together, and that will not appear on their own. Therefore the participant will create a connection between the groups of cues and outcomes, instead of just the individual cues and outcomes. In the experiment, food X was seen in all of the trials, while A and B where both respectively seen in half of the trials. A and B were always seen in combination with X, in other words participants saw either AX or BX in any given trial. These food items were paired together with an outcome that indicated whether an allergic reaction occurred or not.

For each trial the participant had to indicate how likely they thought each of the *three* foods could cause an allergic reaction. They had to indicate this every trial and there were 16 trials per block. The 16 trials always had the same order of cues; the first trial was always AX, the second always BX, the third always BX and so on. This was ordered in such a way that the same trials could never appear more than twice in a row, so for example the trial order AX, BX, BX, BX, AX is not allowed, as BX is seen three times in a row. This order was the same over all participants by design.

Two things differed depending on the block: The first was the type of food that was filled in for A, B and X. In one block a participant would see cheese, pork and blueberries respectively, but in another block they saw strawberries, peanuts and shrimp as A, B and X. There were six of these so called 'food groups', of which each participant saw only three (one in each block).

The other difference between blocks was the outcome condition. There were three different outcome conditions, which we will call 50-50, 75-25 and 100-0 (in the original paper these were 0.00, 0.50 and 1.00 respectively). Each participants was presented with all three conditions once. The condition reflects the probability of AX trials leading to the outcome (an allergic reaction) and that of BX leading to the outcome. In the 50-50 condition, they were both equally likely to lead to an allergic reaction, as 50% of all AX trials would lead to that outcome, and 50% of all BX trials. In one block of 16 trials, this would mean that out of eight AX trials, four would lead to an allergic reaction and four would not, and out of eight BX trials, four would lead to an allergic reaction and four would not. In the 75-25 condition, 75% of all AX trials would lead to an allergic reaction, and only 25% of the BX trials. Lastly in the 100-0 condition all AX trials led to an allergic reaction, but the BX trials would always lead to no allergic reaction. See Table 3.1 for how this would look (with a different outcome, diamond being equal to an allergic reaction, while no diamond is equal to no allergic reaction).

To summarise, each participant saw three blocks of 16 trials each, with each block differing the type of food and the outcome condition, while staying consistent in the cues presented in a certain trial.

In the original experiment participants were asked for a pre-score. This was the score a participant gave the foods seen in that block before seeing any outcomes, thus indicating how likely they thought a food would lead to an allergic outcome in general. We did not simulate this pre-scoring, as we did not give our models any previous knowledge about the foods. As Van Hamme and Wasserman did not report on the pre-scores of the separate foods, we were not able to model after these pre-scores.

2.2 Simulations

Modelling both the Rescorla-Wagner and the Van Hamme-Wasserman model was done in R (R Core Team, 2020), using the package `edl` (van Rij & Hoppe, 2020) for implementation of the EDL formulas and the package `plotfunctions` (van Rij, 2020) for visualising the learning process trial by trial.

The main functions used from the `edl` package were `updateWeights` and `RWlearning`. These functions were originally designed for the Rescorla-Wagner equations and were adapted for the current experiment to run the Van Hamme Wasserman model.

To adapt the original functions, a variable called `etaNeg` is added to `updateWeights`. `EtaNeg` is false by default, which results in the functions doing what they did before and performing Rescorla-Wagner learning. However if `etaNeg` is set to true, the function performs Van Hamme-Wasserman learning. This means that all the cues that are seen up until this moment are in the set of current cues, instead of only the cues that are seen at this

moment. For each cue that is in this set of current cues, the same steps as Rescorla-Wagner learning are followed, as the change in weights between the cue and outcome when a cue is present does not change in the Van Hamme-Wasserman model. The learning rate of each cue that is seen before, but is not in the current set of cues, is multiplied by -1 (the eta is now negative, hence the name `etaNeg`). Then these cues are handled in the same way as the cues that are present. This is as described in Equation 1.2

The only thing changed in the `RWlearning` function is the addition of `etaNeg`, as `RWlearning` calls `updateWeights`. This resulted in changing the name of the `RWlearning` function to `EDLearning` as the function did more than just implement Rescorla-Wagner learning.

With these changes made to the code, the experiment could be modelled. The method of modelling the experiment for both the Rescorla-Wagner learning and the Van Hamme-Wasserman learning was the same, except for the value given to `etaNeg`. What we will describe here are the basics of the implementation, for the code itself see Github *.

A set of cues was created, one for each of the different food conditions (six in total) with an added background cue. A background cue represents all other knowledge that a participant might have in the experiment (Rescorla, 1972). Outcomes for each of the outcome conditions were created as well (e.g. "Allergic" or "Not"). Each modelled participant ran three blocks of 16 trials. After each block, the modelled participant had a saved "memory" of the strength of connection between cues and outcomes. Since the outcomes stay the same across all the blocks, we gave this "memory" as a starting point for the next block.

The experiment is set up such that there are two different outcomes which the model chooses between. This is different from the rating scores that participants had to give. The two outcomes were allergic and not allergic, which reflects a participant choosing between a food definitely not causing an allergic reaction and definitely causing one (so 0 and 8 on the scale from the original experiment). Therefore in the results we will look at these two outcomes in two ways. One of them is the relative connection weight, which is when the connection strength of a cue to not allergic is subtracted from the strength of that cue to allergic. This is also called the relative weight or weight difference, and could be seen as recreating the scale that Van Hamme and Wasserman used. This task would simulate a participant taking their estimate for allergy and their estimate for no allergy and combining them into one measure. The other way in which we will look at these outcomes is in the individual connection weight to allergic and not allergic. Looking at both of these individually will give an idea of what happens over the course of the experiment, which, if looked at both would give a better idea of how the predictiveness of cues to these two outcomes changes over the experiment.

*https://github.com/SannePoelstra/MasterProject in the folder *Experiment*

2.3 Simulation results

In this result section, for reasons of clarity, we will be presenting the results from condition 75-25 only. Conditions 50-50 and 100-0 (and averages over all conditions) can be found in Appendix B, there are no large differences between the results discussed here and those shown in the Appendix, unless it is specifically mentioned in this section.

Figure 2.1 shows the original experiment results from Van Hamme and Wasserman's 1994 paper on the left. Here the average causal rating scores over all participants over trials of condition 75-25 are displayed. On the y-axis we see the causal rating score (from 0 to 8) and on the x-axis the 16 trials. In this plot the pre-scoring is also shown, the score given to the foods at the beginning of the block, on the left of the grey line. As can be seen, all three pre-scored fall around a score of 4, which means that each of these foods could *possibly* lead to an allergic reaction.

In this left plot we see that participants learn that cue A is a better predictor of an allergic reaction than B, as the causal rating scores for A gradually increase, and the scores for B decline over the course of the block. The fluctuation within the ratings for any of these cues is due to that cue either appearing or not appearing on a certain trial and what the paired outcome to that cue was.

In the middle of Figure 2.1 we see the predictions that the Rescorla-Wagner model makes. Here the plot shows the average connection weight for all simulated participants over trials. As the models did not give a rating, the y-axis is now represents the connection weight from allergic minus those to not allergic (weight difference). The x-axis is still the same, showing the 16 trials. As we did not let the models give a pre-score, we only see the results of the 16 trials. In this middle graph we see that the model also finds that cue A is more predictive (has a higher connection weight) of an allergic reaction over trials, while B gets less predictive of an allergic reaction over trials.

On the right of Figure 2.1, the predictions of the Van Hamme-Wasserman model are shown. The y-and x-axis are the same as in the middle graph, connection weight difference and trials. Over trials cue A is more predictive of an allergic reaction, while B becomes less predictive.

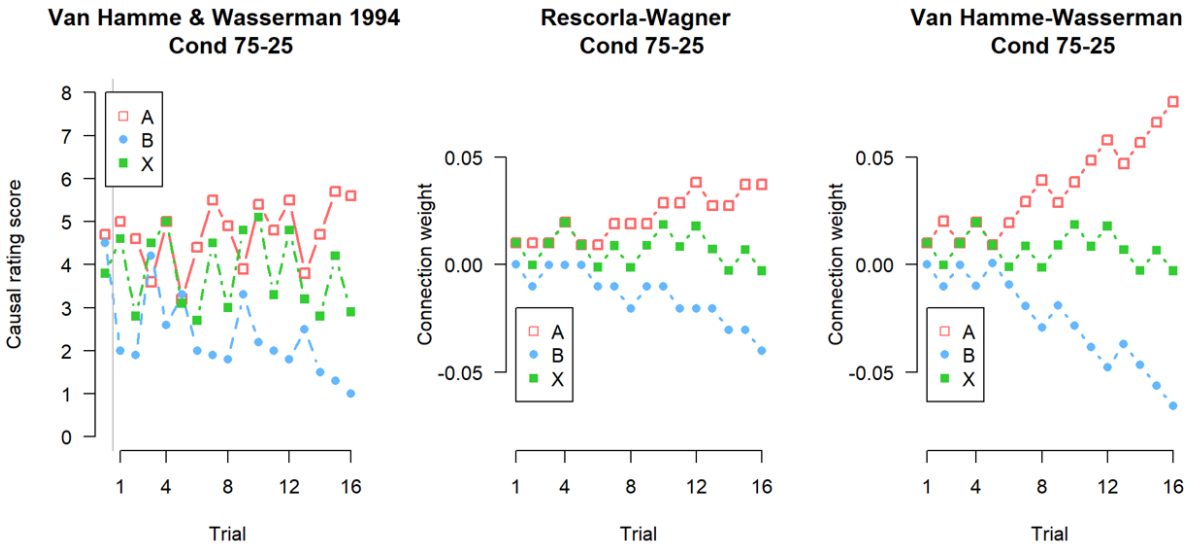

Figure 2.1: Left: Data from the original paper, average causal rating over all participants per trial; Middle: Rescorla-Wagner model, average connection weight of allergic minus not allergic over all simulated participants per trial; Right: Van Hamme-Wasserman model, average connection weight of allergic minus not allergic over all simulated participants per trial.

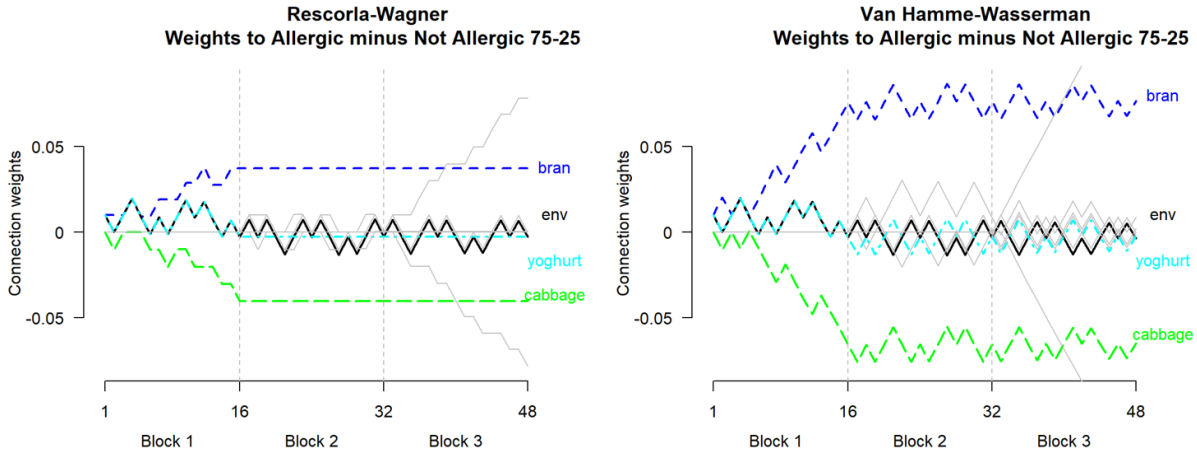

Figure 2.2: Weights to allergic minus the weights to not allergic for each of the foods asked. Left: Rescorla-Wagner model. Right: Van Hamme-Wasserman model.

While the Van Hamme-Wasserman model shows more spread out connection weights (A increases further and B decreases further than in the Rescorla-Wagner plot), they both make a very similar prediction. It seems that the average over trials does not make for a clear difference in prediction between the Rescorla-Wagner model's behaviour and that of the Van Hamme-Wasserman model. It would therefore be interesting to investigate the behaviour of the connection weights over the whole experiment, as this is where we would expect the models to differ more, since the models handle absent cues differently and the cues are only seen within their respective blocks.

In Figure 2.2 we see the weights to allergic minus those to not allergic for the Rescorla-Wagner model on the left and the Van Hamme-Wasserman model on the right. For ease of interpretation, the cues from the first block are shown in colour, while the others are greyed out. The division between the three blocks is indicated with a light grey, vertical dotted line. As for both allergic and not allergic the weights stay constant in the Rescorla-Wagner model, the lines in this graph also stay constant. However in the Van Hamme-Wasserman function the weights to allergic and not allergic will fluctuate, as can be seen in the graph for the cues after block 1.

Although the strength of activation (i.e. model expectation) is different, qualitatively the two models make the same predictions.

As mentioned before there is a difference between looking at the relative connection weights and the individual weight to either of the two outcomes. As we saw in Figure 2.1, both the models showed a similar pattern to that of Van Hamme and Wasserman's original experiment, however there is less fluctuation within the cues themselves. This might be because relative connection weight is more similar to a participant choosing between a score of 0 and 8, instead of making use of the whole rating scale. It would therefore be interesting to investigate the behaviour of the connection weights over the whole experiment, to both allergic and not allergic.

The top of Figure 2.3 shows the weights to allergic for the Rescorla-Wagner model on the left and the Van Hamme-Wasserman model on the right. Please note that the scales of these two graphs is not the same, as the right plot displays a bigger scale than the one on the left. This was done to present both results clearly, which could not be done well when the scales were kept the same. The division between the three different blocks is indicated with a vertical grey line and the cues from the first block are highlighted. For now we will just look at blocks 1 to 3, which we will call the training phase. Cabbage is a B cue, bran an A cue and yogurt an X cue.

In the Rescorla-Wagner model the weight of a cue that is no longer encountered stays at the same weight, while the weight of a cue that is no longer encountered in the Van Hamme-Wasserman model decreases in weight over trials. This is in line with the Equations these models are based on, as the first line from Equation 1.1 (Rescorla-Wagner) states the weights of absent cues are not updated. The first two lines of Equation 1.2 (Van Hamme-Wasserman) state that an absent cue and absent outcome will lead to a decrease in weight and an absent cue and present outcome will lead to an increase in weight. The fluctuation of this latter equation can be seen in the fluctuation in the decreasing line of Figure 2.3, however as there is now more cue competition, the activation (sum of all weights) does not increase as steeply when as in the first block, and thus the weight decreases over all.

The bottom of Figure 2.3 shows the same as the top, but now for the weight to not allergic. We can see that compared to the weight to allergic, the lines of the B cue (cabbage) and the A cue (bran) are now switched around, as the B cue leads to the outcome not allergic in 75% of the BX trials.

In this Figure we see a similar effect to that of the previous one, where the weights in the Rescorla-Wagner model stay constant when a cue is no longer encountered, while the weights in the Van Hamme-Wasserman model decrease. This similar effect might indeed indicate that allergic and not allergic should not be seen as two separate and individual outcomes, but more as this kind of concept that works together, which would be better captured in the scale that Van Hamme and Wasserman used.

Van Hamme and Wasserman only ask for cues within their respective block, therefore the decline in predictability of a cue is not something that they would have captured.

We propose that adding a test phase to the end of the experiment, where cues from the first and last block are asked again, will be able to create differing predictions for these two models. In this test phase participants were asked only two cues and their relation to an allergic reaction (so no longer not allergic). We simulated this test phase, with an extra inclusion of a new cue not asked before in the training phase, of which the results will be discussed below.

In Figure 2.3 the test phase can be seen, after the three blocks of the training phase. The blue arrow in the top and bottom of the Figure indicates the weight of the new stimulus that is only seen in the test phase.

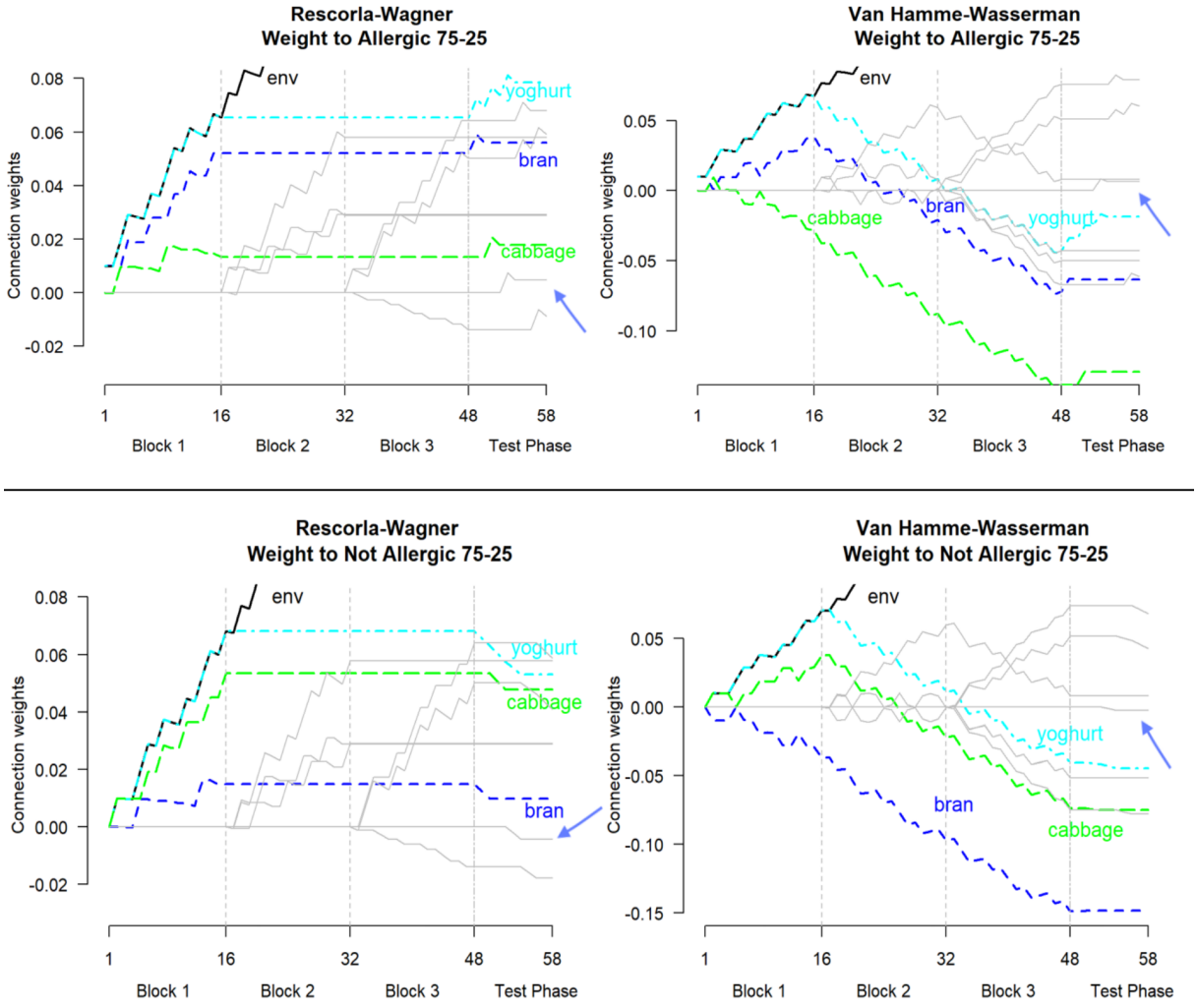

Figure 2.3: Weights to: Top: allergic, Bottom: not allergic, with test phase, block 1 is condition 75-25. The arrows indicate the novel cue that is only introduced in the test phase. Left: Rescorla-Wagner model. Right: Van Hamme-Wasserman model.

In the Rescorla-Wagner model in the top part of Figure 2.3, the test phase cues are still at approximately the same weight level as in the training phase. This is the case for both the first block and the third block. When comparing the previously seen cues to the weight of the new cue, we see that they are more predictive (have a higher connection weight) of an allergic reaction than the new cue.

If we look at the Van Hamme-Wasserman model in the top part of Figure 2.3 we can see that the cues from the first block have decreased considerably in the test phase when compared to their original block. However cues from the third block are still relatively similar to their original block in the test phase. There is thus a big difference in pattern between recently asked cues and cues that were asked in an earlier phase. Cues from the first block are also worse predictors of an allergic reaction than a new cue, as they all have lower connection weights when compared to the new cue. However the cues from the third block are all either higher or the same.

In the bottom part of Figure 2.3 we see similar results. However as now we have a present cue, but an absent outcome (we did not ask for the foods relation to not allergic in the test phase), we can see the weights decrease in the test phase for both models (as is in line with the second line of Equation 1.1 and the third of Equation 1.2). It is important to note that here too in the Rescorla-Wagner model the cues that are seen before have higher connection weights (and are thus more predictive) of no allergic reaction than a new one is. So what we see in the allergic plots (the new cue being lower than the already seen ones), does not mean that this new cue is a better predictor of a lack of an allergic reaction, which we see in this plot. It is therefore also important to add a new outcome to the test phase, to which to compare the two pre existing outcomes to.

2.3.1 Discussion

Van Hamme and Wasserman conceptualised the issue as learning in the absence of cues. However the simulations show a slightly more complex story, namely the models show no difference in relative activation (Figure 2.1 and Figure 2.2). The only difference that emerges is the weight of the cues that are no longer encountered to allergic or not allergic, as could be seen in Figure 2.3. In the Rescorla-Wagner model these weights stay constant, but in the Van Hamme-Wasserman model they decline. A way to test which of these two predictions holds, is to introduce a test at the end of the experiment. One with both new and old stimuli, to compare if the activation of the old stimuli has decreased compared to when they were seen and to see if this is then lower than a new stimulus that has not been seen before.

We expect that if the Van Hamme-Wasserman model is correct, there should be a big difference between the scores given to the cues in block one and the scores given to those same cues in the test phase. This difference in score then should be smaller between the cues asked in the third block of the training phase and the test phase. This will be the case for both an allergic reaction *and* to no allergic reaction. We expect the score for a new cue to be higher than the scores for the cues asked in the first block of the training phase.

If the Rescorla-Wagner model is correct, then there should be no or a small difference between the scores given in the first block, and those given in the test phase. The score difference between the first block and the test phase and the third block and the test phase should be small or non-existent, as the cues from the first block do not decrease in weight when they are no longer seen, so there should be almost no difference in weight between a cue from the first and the last block. When compared to new cue we expect the cues to be more predictive of an allergic reaction (and no allergic reaction).

Modelling is a very useful tool in creating predictions, as we can create more precise predictions then just verbal ones. We can also learn more about our experiment design and if the current design set up would actually test what we want to test.

Modelling is almost always a simplification however, so there are some things to take into account. The first one is that we assume here that a higher weight is a higher predictability and will thus lead to a higher score. However there might still be some explicit interference happening which makes this translation step more complicated.

The fact that the relative activation did not show a difference in prediction between the two models might be because in this modelling the outcomes are more similar to choosing between a score of 0 (definitely not causing an allergic reaction) and 8 (definitely causing one), instead of rating according to a scale. It is also important that, while we do look at allergic and not allergic separately, it is difficult to fully pull them apart.

As we said, modelling will almost always be a simplification of reality, so therefore it is important to test these predictions with an experiment. In the next sections we will describe the three experiments we did.

Each of these experiments will have their own methods, results and discussion sections. At the very end of the thesis we will have a general discussion in which we will discuss the results of these experiments combined.

3 Experiment 1

The aim of this experiment is to replicate the findings reported by Van Hamme and Wasserman and to see if their findings generalise to other experimental stimuli, as well as testing the hypotheses obtained through computational simulations.

We will partially replicate the original experiment done by Van Hamme and Wasserman in their 1994 paper, and test some additional predictions from the modelling without changing the training design. We do this by adding a test phase after the training design, as we wish to test the predictions made by the model (see the previous section for the exact predictions). In this test phase we will also present different stimuli compared to those seen before and re-frame the outcome.

3.1 Methods

3.1.1 Participants

The participants were selected via the online platform Prolific (Palan & Schitter, 2018). This is similar to *Amazon Mechanical Turks*, but created with a bigger focus on the scientific community.

No demographic data was asked of the participants, however they were selected based on the following criteria: They had to be between 18 and 25 years old, as the original Van Hamme and Wasserman experiment used undergraduates and we wanted to match that group. They also had to be fluent in English, but not from a specific country. In total 86 participants took the experiment, of which one timed out, 19 returned their submission themselves (and thus indicated that they no longer wanted to participate), and six were rejected. Of those six, two participants did not consent on the consent form and were automatically rejected [†], the other four handed in answers that were deemed not serious. This was either because they filled in the same number everywhere (or variants such as 1,2,3 then 4,5,6 etc.), more than half of the values were not filled in (NA's), or they only spend half a minute in the experiment before finishing. This meant that we were left with 60 participants in total. The participants received a monetary reward of £2.50 for the experiment and they were told it would take 20 minutes (average completion time of 18 minutes).

3.1.2 Materials/Stimuli

We changed the stimuli compared to the 1994 experiment. As mentioned before, there is already

[†]Prolific does state that it is better to send participants a message and let them return their own submission instead of the researcher rejecting them if this happens, this was done in later experiments.

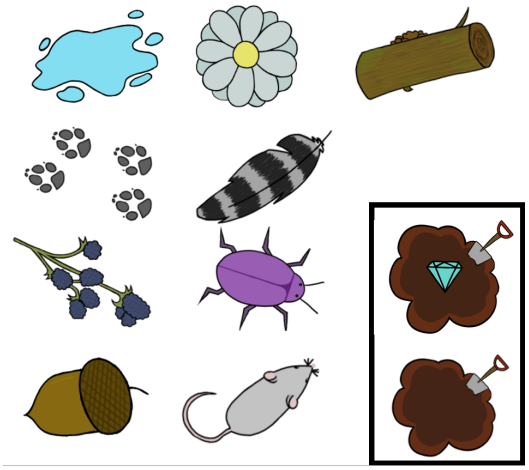

Figure 3.1: Cues and outcomes shown in the training phase of Experiment 1. Outcomes are in the black box.

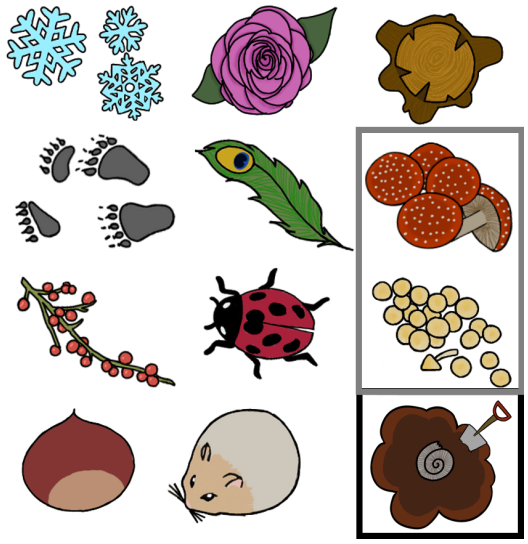

Figure 3.2: Cues and outcomes shown in the test phase of Experiment 1. Outcomes are in the black box, the new cue plus its alternative in the grey box.

a pre-existing association between food items and allergic reactions. Therefore we wanted to create new neutral cues that did not have any previous connections to the outcome. We also wanted to avoid participants forming categories based on cues, therefore we opted for natural stimuli instead of abstract ones, as people often classify abstract stimuli into categories based on traits, which could influence our results (Lassaline et al., 1992).

The stimuli in this current experiment are all objects that one could find on the forest floor. The cues and outcomes shown in the training phase can be seen in Figure 3.1. Here we see a puddle, footprints (wolf), blackberries, an acorn, a flower, a feather, a beetle, a mouse and a log. The two outcomes are a diamond (which is the equivalent to an allergic reaction) or an empty hole/no diamond (no allergic reaction).

In the test phase participants saw both old cues and outcomes, as well as new ones, which can be seen in Figure 3.2. Snowflakes, footsteps (bear), red berries, a chestnut, a rose, a peacock feather, a ladybug, a hamster and a stump, all corresponding to the old cues seen in the same place in Figure 3.1. As an addition to the old cues, toadstools are also introduced as a new cue in the test phase. The alternative to this new cue is right below it, small brown mushrooms. The alternative outcome is a fossil.

The stimuli were all drawn in Krita 4.3.0 on a Huion Camvas 16 Pro tablet. The images were 211 x 152 px to fit on the screen. The stimuli were created especially for this experiment by the author.

3.1.3 Experimental Design

As can be seen in Figure 3.1, there were a total of nine cues. These cues were randomised over participants. This meant that one participant might see a mouse for cue A in block 1, while another might see a log for that same A cue. This was done to reduce the influence of individual connections between cues and outcomes.

Participants had to rate the likelihood of the cues leading to a diamond on a scale from 1 to 9. This was changed from 0 to 8 in the original experiment, as on a keyboard the 0 is placed on the right of the other numbers and might therefore be less logical to rate. The number pad (numpad) could not be used.

The training phase consisted of three blocks of 16 trials. The order of the cues in these trials was always the same (as in the original experiment), but the outcome depended on the condition. Which order of conditions participants saw depended on their subject number, through which they were split into six groups (as there are three conditions, so six unique ways to display them).

The order of the cues presented and their respective outcomes can be seen in Table 3.1, this is the same order as Van Hamme and Wasserman used in their experiment. Just as in the original experiment, there could be no more than three of the same compound cues in a row (so AX, BX, BX, BX, AX would not be possible).

The test phase consisted of three different parts. The cues and outcomes of the first part can be seen in Table 3.2. All the cues with a 1 at the end are the same cues seen in the first block of the training phase, while those with a 3 at the end where the same cues as seen in the last block. C is a completely new cue that participants had never seen before.

Each of these test phase parts had a different purpose. The focus of the first part was to ask participants about cues that they had already seen and score them. The cues were given in the same order as they had appeared in their original training phase blocks, so they would see the cues from block 1 first, and after that cues from block 3. Block 2 from the training phase was omitted, as we wanted to see the effect of the first and the last presented cues, as we expected the biggest difference to appear between these two. In the test phase participants did not see an outcome in the same location as in the training phase, instead they would see a question asking the likelihood of a certain outcome (see Figure 3.5). This meant that participants were asked about a certain outcome, but they never got the confirmation that they had in the training phase, on whether or not these cues would lead to that outcome, and therefore participants should not learn in the test phase. The outcome they saw was either a diamond, which they had seen before, or a fossil, which was a completely new outcome.

The second part of the test phase had two separate goals. The first was to increase statistical power, as if there was no difference between the scores of the first and second part of the test phase, these two could be collapsed into one data set for the analysis. The second goal was to check if learning took place in the test phase. If no learning took place, then the scores between the first and second part should not differ. This part used the same stimuli as the first part of the test phase, but they were now shown in random order.

The last part of the test phase consisted of completely new cues that were related to the original ones. For example, if a participant had seen a mouse for their A1 cue, then they would see a hamster in this part of the test phase. While in the previous two parts we only showed a diamond and a fossil outcome, here the no diamond outcome was reintroduced. The combination of cues and

Trial	Cues	50-50	75-25	100-0
1	A X	Diamond	Diamond	Diamond
2	B X	No Diamond	No Diamond	No Diamond
3	B X	Diamond	Diamond	No Diamond
4	A X	No Diamond	Diamond	Diamond
5	A X	No Diamond	No Diamond	Diamond
6	B X	Diamond	No Diamond	No Diamond
7	A X	Diamond	Diamond	Diamond
8	B X	No Diamond	No Diamond	No Diamond
9	B X	Diamond	Diamond	No Diamond
10	A X	No Diamond	Diamond	Diamond
11	B X	No Diamond	No Diamond	No Diamond
12	A X	Diamond	Diamond	Diamond
13	A X	No Diamond	No Diamond	Diamond
14	B X	Diamond	No Diamond	No Diamond
15	A X	Diamond	Diamond	Diamond
16	B X	No Diamond	No Diamond	No Diamond

Table 3.1: The cues shown in each trial, plus the outcome to those cues based on the conditions (50-50, 75-25 and 100-0).

Cue 1	Cue 2	Outcome
A1	X1	Diamond
A1	X1	Fossil
B1	X1	Diamond
B1	X1	Fossil
C	X1	Diamond
C	X1	Fossil
A3	X3	Diamond
A3	X3	Fossil
B3	X3	Diamond
B3	X3	Fossil

Table 3.2: Test phase parts 1 and 2 (order randomised for part 2). Where A1, B1 and X1 are the cues seen in training phase block one; A3, B3 and X3 those seen in block 3; and C is a completely new cue.

outcomes of this third part of the test phase can be seen in Table 3.3. The purpose of this last part was to gather data for potential future research. We wanted to know if participants would recognise that these new cues belonged to the same 'category' and would score them similarly, meaning that participants did not just learn connections between an object and an outcome, but also learned a connection of the properties of that object to the outcome. As this last part of the test phase works with new cues that were not seen before, we cannot collapse the results from this part into one data set together with the results of the first and second part of the test phase.

In total the test phase consisted of 35 trials (10+10+15), therefore participants saw a total of 83 trials in the whole experiment, plus four practice trials (that we did not analyse).

Cue 1	Cue 2	Outcome
A1_alt	X1_alt	Diamond
A1_alt	X1_alt	Fossil
A1_alt	X1_alt	No Diamond
B1_alt	X1_alt	Diamond
B1_alt	X1_alt	Fossil
B1_alt	X1_alt	No Diamond
C_alt	X1_alt	Diamond
C_alt	X1_alt	Fossil
C_alt	X1_alt	No Diamond
A3_alt	X3_alt	Diamond
A3_alt	X3_alt	Fossil
A3_alt	X3_alt	No Diamond
B3_alt	X3_alt	Diamond
B3_alt	X3_alt	Fossil
B3_alt	X3_alt	No Diamond

Table 3.3: Test phase part 3. Where A1_alt, B1_alt and X1_alt are the alternative cues to those seen in training phase block 1; A3_alt, B3_alt and X3_alt alternatives to those seen in block 3; and C_alt is the alternative to the completely new cue.

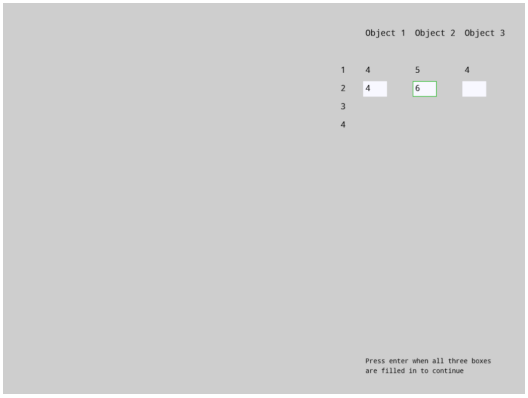

Figure 3.3: Example of a screen that a participant saw in practice of Experiment 1. There were no example cues nor outcomes.

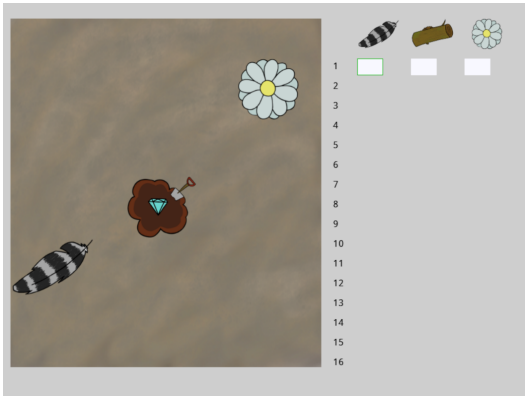

Figure 3.4: Example of a screen seen by a participant in the training phase of Experiment 1.

3.1.4 Procedure

All the experiments mentioned in this thesis were created in OpenSesame version 3.3.7 (Mathôt et al., 2012), with JavaScript inline code. They were run on a Jatos server and presented to participants via Prolific.

The participants could see the description and requirements of the experiment on Prolific. Once they chose to participate, they saw a consent form. If they did not consent, the experiment would stop automatically. They could also withdraw (return) their experiment results at any given moment.

Participants were given an explanation of the goal of the experiment. In short, they were space explorers on a foreign planet looking for treasure. To find treasure they had to pay attention to their surroundings. They would be relocated to a different part of the planet after 16 days and they would move twice. Each "day" corresponded to one trial where two cues and one outcome were seen. They were not told about the test phase, as we wanted them to learn naturally over the course of the training phase.

Each of the text screens in the experiment would only let a participant progress to the next screen if they clicked a button to confirm that they wanted to move to the next screen.

Once a participant confirmed they understood the instructions, they could practice with the rating system. In Figure 3.3 an example of such a practice round can be seen. Participants could move between boxes with the arrow keys and type in any number from 1 to 9 on their keyboard (but not the numpad) in the boxes. The green line around the box would indicate which box they had selected. Once all three values were filled in they could press enter to go to the next day. Notice that in the practice trials we do not show pictures yet, as we do not want participants to start learning already.

If a participant finished the practice run they were notified of the fact that now the real experiment would start.

In the training phase (Figure 3.4) a participant would see two cues (out of the three that were assigned to that block) on screen at semi random positions each trial. In the middle appeared either a diamond or an empty hole. The ratings of the previous trials were shown above the rating boxes of the current trial. First the cues were shown for 2000 ms, then the cues together with the outcome for 1000 ms, after which participants had 13000 ms to fill in their answer. Participants could fill in their ratings in the same manner as with the practice trials. If they finished before the time ran out, they could press enter to continue. If they did not fill in anything before the 13000 ms had passed, the cell would automatically be filled in with an "X".

After participants finished a block, a bit of text would appear to tell them that they would now go to the next part of the planet, where they would find different clues (cues).

Once all training phase blocks were finished, participants were informed that they had to make predictions for the future, which was the test phase of the experiment. This meant that they would not see the outcome appear, but there would just be a question asking "How likely will these objects lead to the result in the picture below?" (Figure 3.5). They would only have to give a rating for the two cues that they saw on screen in any given trial.

The average completion time of the experiment was 18 minutes. Ranging from about 10 minutes to an hour. Even though the trials themselves were restricted to a time limit, the explanation screens were not, therefore the experiment could take longer than one would expect looking at the time limit per trial.

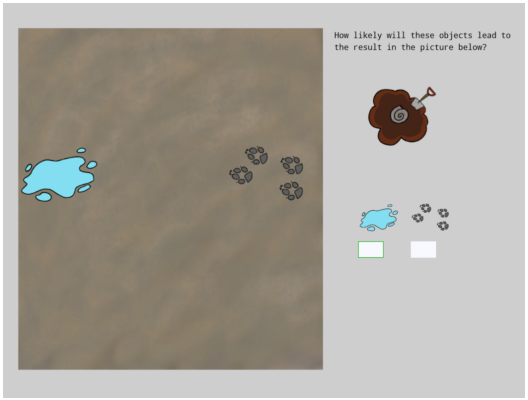

Figure 3.5: Example of a screen seen by a participant in the test phase of Experiment 1.

3.2 Results

The results are split into two sections. The first is the training phase, where we compare the results of the current experiment to those of Van Hamme and Wasserman. The second section concerns the test phase, where we can see whether our model predictions hold. In this results section we will only discuss results that answer our main research question. Analyses that are done outside of that are discussed in Appendix A.

3.2.1 Training Phase

We will first compare the results of Van Hamme and Wasserman's paper to our own results. Figure 3.6 contains multiple plots, comparing our results to that of Van Hamme and Wasserman. For all of these plots, the data from Van Hamme and Wasserman is on the left, and that of the current experiment is on the right. As error bars were not displayed in the original paper, we also did not display them. However the values for the standard error of the mean for all these points, plus plots including the standard error of the mean can be found in the analysis files for all experiments on Github[‡] for those that are interested.

We will first discuss the average causal rating scores per condition, which are shown in the plot in the top left of Figure 3.6. In these plots the x-axis displays the three conditions (50-50, 75-25 and 100-0) and on the y-axis the causal rating score can be seen. Note that this causal rating score ranges from 0 to 8 in the original experiment, but in our current experiment participants rated from 1 to 9.

On the left we see that the higher the chance that cue A leads to an allergic reaction (in our experiment a diamond), the higher the average score of cue A, while cues B and X decrease in score. In our experiment, while cue A shows a very slight increase and B and X a slight decrease, all three cues seem to average between 3 to 5, regardless of condition. This could indicate that participants did not understand what they had to do. Cue B is scored consistently lower than the other two cues however, this is similar to the results of the original paper.

It is also of interest to look at the individual conditions and investigate what happens over trials, as we still might see a similar pattern over trials emerge as in the Van Hamme and Wasserman results.

In the top right of Figure 3.6 we see the average causal rating score for all participants over trials for condition 50-50. The grey line in the left plot separates the pre-scoring (which our participants did not do), from the rest of the trial data.

While we do see a similar pattern between the two experiments, the fluctuation of the ratings is smaller in the right plot. For example, Van Hamme and Wasserman's score for cue X fluctuates considerably, depending on the cue it is shown with. However on the right, we can see that our score for X stays between and average rating of 3 and 5.

The bottom left of Figure 3.6 shows the average causal rating score for all participants over trials for condition 75-25. This is the same plot that we looked at when comparing out modelling results to the original paper. In the left plot we can see participants started to learn that over trials, cue B is a worse predictor of an allergic reaction than the other two cues, while A is a better one. In the right plot we do see that B is scored slightly lower than the other two cues, but cue A and X are still very much entangled.

Lastly the bottom right part of Figure 3.6 shows the average causal rating score for all participants over trials for condition 100-0. In the left plot we now see a very clear distinction between cues. Cue A is scored very high, as it always predicts an allergic reaction, while B is learned to be not very predictive (or very predictive of a lack of an allergic reaction) and thus gets a lower score. As X is shown with both cues A and B, this cue's score also drops, as it is not as predictive of an allergic reaction as cue A is. In our experiment data on the right, we do see a lower score for B. However we do not get the distinction between cues A and X that Van Hamme and Wasserman found. We also do not have a very wide range of ratings, with the ratings staying between an average of 3 to 6.

The fact that these scores centre more around the middle of the rating scale than those of Van Hamme and Wasserman, could be due a couple of reasons. Participants might for example just have answered a rating of 5 (meaning possibly leading to

[‡]https://github.com/SannePoelstra/MasterProject in the folder *Results*

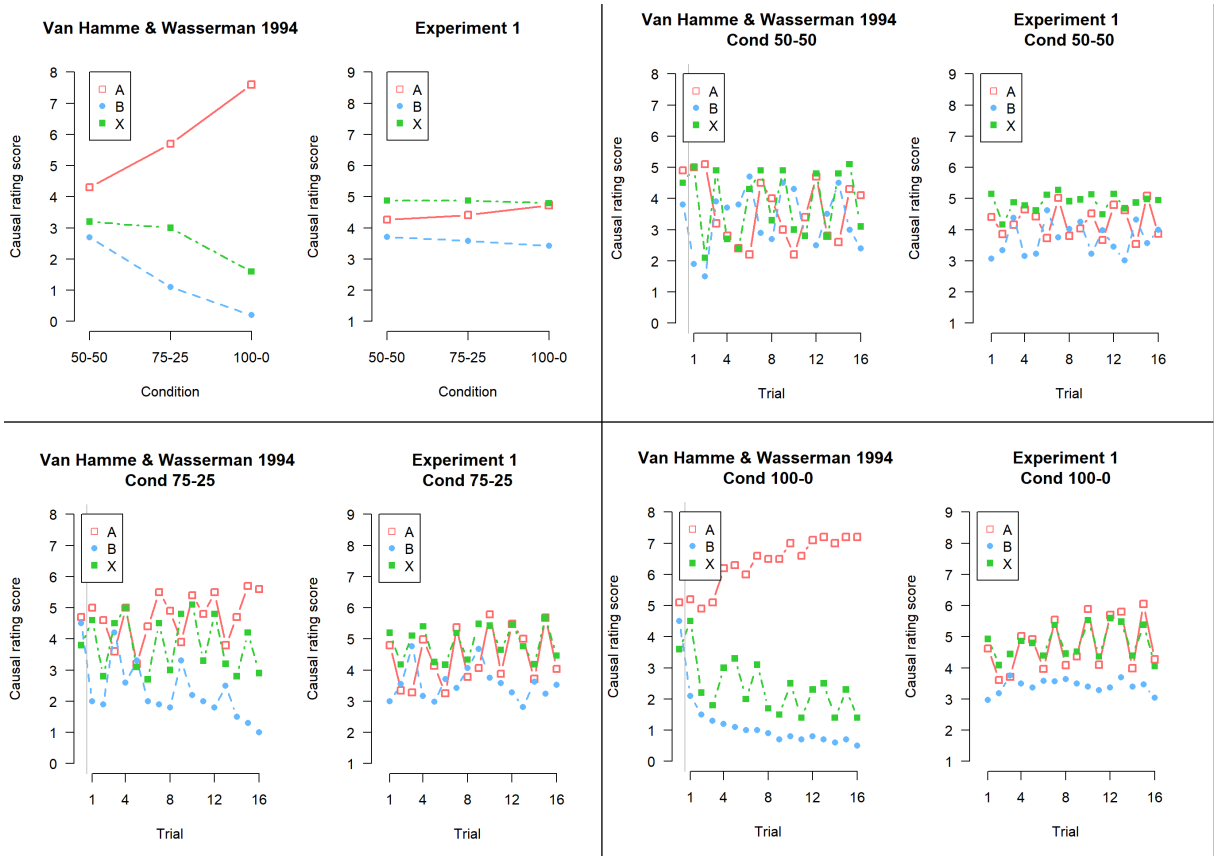

Figure 3.6: Average scores of cues A, B and X of Van Hamme and Wasserman's 1994 paper (left) and that of Experiment 1 (right). Top left: average causal rating scores over conditions, top right: average causal rating scores over trials for condition 50-50, bottom left: average causal rating scores over trials for condition 75-25, bottom right: average causal rating scores over trials for condition 100-0. The grey line vertical line in the plots over trials separates the scores for the pre-scoring from the rest of the trials.

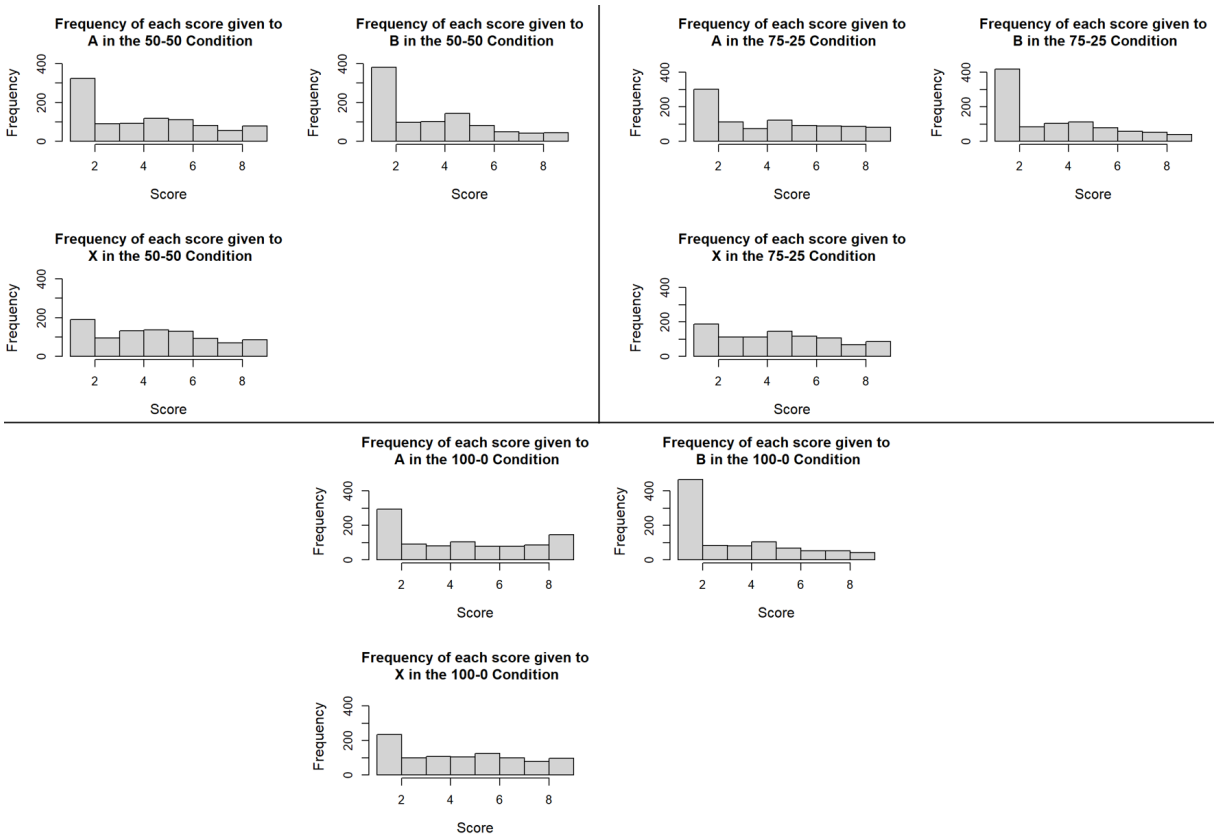

Figure 3.7: Frequency of each of the scores given to cues A, B and X of Experiment 1. Top left: condition 50-50, top right: condition 75-25, bottom: condition 100-0.

a diamond) more often, or they could have scored on the two extremes (1 and 9). Therefore we will now look at the frequencies of the scores given to each of the cues per condition.

In Figure 3.7 the frequencies of the causal rating scores for cues A, B and X are shown. In the top left we see condition 50-50, in the top right condition 75-25 and on the bottom condition 100-0. We will not discuss all of the plots in this Figure, but we will highlight those that are of interest.

In condition 50-50 both AX and BX led to a diamond in half their trials, and to no diamond in the other half. One would then expect that the scores given to cues A and B are either quite varied, or that they are given a score of 5 (possibly) most often, as it is quite unsure if they led to a diamond or not. However in the top left of Figure 3.7 we can see that both cues A and B were scored a 1 (definitely not leading to a diamond) most often. In condition 100-0, where AX always led to diamond outcome while BX always led to a no diamond outcome, one would expect cue A to be scored a 9 (definitely leads to a diamond) most often and B to be scored a 1 most often. While B is scored in line with what we would expect, we see that cue A is still scored a 1 most frequently. It seems then that participants did not fully understand the scoring.

As the scores differ less per condition as those of Van Hamme and Wasserman, and the participants did not seem to fully understand the scoring, we will now look at whether or not participants scored a diamond trial higher than a no diamond trial. If there was a difference between these two scores, it would suggest that participants at least understood how they had to score on a trial-to-trial basis.

We compared the scores on the no diamond trials to those on the diamond trials with two-sided, paired t-tests. As we did multiple t-tests over the same data-set, we adjusted the p-values according to the Bonferroni-Holm method (Holm, 1979) to reduce the probability of obtaining Type I errors. These adjusted p-values are also the ones that we will report. The averages of the scores for the diamond and no diamond trials can be seen in Figure 3.8. This Figure depicts the scores for the cues in all the trials, so trials in which the cue appeared, but also those in which the cue was absent.

In condition 50-50 cues A ($t(59) = 2.49$, $p = .08$), B ($t(59) = 1.73$, $p = .27$) and X ($t(59) = 2.28$, $p = .10$) were not scored significantly different in a diamond trial versus a no diamond trial. For conditions 75-25 and 100-0 both cues A ($t(59) = 5.16$, $p < .001$; $t(59) = 4.87$, $p < .001$) and X ($t(59) = 4.31$, $p < .001$; $t(59) = 3.05$, $p = .02$) scored

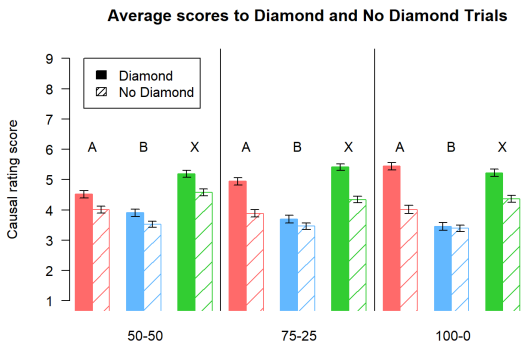

Figure 3.8: The average scores to each condition for each cue on diamond (solid colour) trials and no diamond (hatched) trials for Experiment 1. The error bars represent the standard error.

higher on diamond trials than on no diamond trials. However, cue B ($t(59) = 0.81$, $p = .84$; $t(59) = 0.25$, $p = .84$) did not differ for either of these two conditions.

So overall for all conditions except 50-50, cues A and X scored higher on the diamond trials than the no diamond trials. However for cues A and X in the 50-50 condition and for all B cues regardless of condition, there was no significant difference found. Therefore it seems as though, on a trial to trial basis participants might not have fully understood what exactly they had to score to.

3.2.2 Test Phase

The test phase was introduced to test the predictions obtained through our computational simulations, which we will repeat here for clarity before discussing the results.

In Figure 2.3 we saw that for the weights to allergic, the Rescorla-Wagner model predicted that cues that were asked in the first block still had the same connection weight when they were asked in our modelled test phase. The same can be seen for the connection weights of the cues from the last block in the training phase that participants saw. We assume here that this connection weight is related to the scores, so that the scores will also stay the same.

In Figure 2.3 we could also see that the Van Hamme-Wasserman model predicts a decrease in connection weights to allergic for cues from the first block when they are seen again in the test phase. This decrease is not seen for cues from the last block, as those have not been able to decrease in weight yet.

We also looked at the connection weights related to a new cue introduced in the test phase. The Rescorla-Wagner model predicts that cues that have been seen before in the training phase, have a higher connection weight to allergic than a cue that has not been seen before. In the Van Hamme-Wasserman model, the cues from the first block have a lower connection weight than the new test phase cue, but cues from the third block are scored higher.

We will check which of these predictions hold by looking at the first two parts of our test phase. As there was no difference found between the scores in these two parts, we combined them into one data set. As we perform multiple t-tests over the same data-set, we use and report on the adjusted p-values according to the Bonferroni-Holm method for this whole result section.

Due to the fact that the cues behaved differently depending on condition, for example, cue A predicted a diamond more often in 100-0 than in 50-50 and would thus be scored differently depending on the condition, we will look at both the effect of the cues overall and the effect of the cues per condition.

As mentioned in the beginning of the results section, we will only report on the results that help answer our research question and compare our research to that of Van Hamme and Wasserman in this section of the thesis. The results of the third part of the test phase and the comparison between part 1 and 2 can be found in Appendix A.

Difference in scores training and test phase: Table 3.4 contains the results of the Wilcoxon tests between the scores to the cues in the training phase and the test phase. We used a Wilcoxon test instead of a t-test, as the difference in data was not normally distributed. The squares that are coloured orange are cues for which the scores decreased from training to test phase, the red colour are cues that were significant but with the Holm correction are no longer and the blank squares were not significantly different.

If we look at the conditions overall (in the column All), cue A seen in the first block was scored lower in the test phase than in the training phase. However when looking at the conditions individually there was no significant difference between the scoring in the test and training phase for any of the cues from the first or the third block. This would support the Rescorla-Wagner model, as this model predicted that there would be no significant difference between the scores in the training and in the test phase. While there was one cue that decreased in score, this was not found when looking at the separate conditions, therefore the support seems to be more in line with Rescorla-Wagner, than with Van Hamme-Wasserman.

	All	50-50	75-25	100-0
A1	$V = 418$, $p = .004$	$V = 79$, $p = 1$	$V = 31$, $p = .28$	$V = 43$, $p = .32$
B1	$V = 899$, $p = .058$	$V = 130$, $p = 1$	$V = 59$, $p = 1$	$V = 130$, $p = .22$
X1	$V = 992$, $p = .35$	$V = 129$, $p = 1$	$V = 110$, $p = 1$	$V = 125$, $p = 1$
A3	$V = 919$, $p = .43$	$V = 90$, $p = 1$	$V = 145$, $p = 1$	$V = 82$, $p = 1$
B3	$V = 566$, $p = .65$	$V = 19$, $p = 1$	$V = 61$, $p = 1$	$V = 158$, $p = 1$
X3	$V = 1024$, $p = .26$	$V = 28$, $p = 1$	$V = 115$, $p = 1$	$V = 221$, $p = .22$

Table 3.4: Differences in scores to cues in the training phase and test phase of Experiment 1. Red: used to be significant but no longer with the Holm adjustment, Orange: test phase scored lower than training phase, Blank: no significant difference in scores.

	All	50-50	75-25	100-0
A1	$V = 601$, $p = .09$	$V = 914$, $p = .37$	$V = 891$, $p = .73$	$V = 939$, $p = .46$
B1	$V = 442$, $p = .007$	$V = 809$, $p = .09$	$V = 820$, $p = .60$	$V = 820$, $p = .10$
X1	$V = 338$, $p < .001$	$V = 622$, $p = .003$	$V = 640$, $p = .06$	$V = 698$, $p = .01$
A3	$V = 243$, $p < .001$	$V = 320$, $p < .001$	$V = 608$, $p = .002$	$V = 772$, $p = .005$
B3	$V = 671$, $p = .16$	$V = 1080$, $p = .70$	$V = 1271$, $p = .95$	$V = 1558$, $p = .95$
X3	$V = 349$, $p = .003$	$V = 665$, $p = .60$	$V = 735$, $p = .03$	$V = 857$, $p = .02$

Table 3.5: Differences in scores to cues in the test phase and scores to the new cue C for Experiment 1. Red: used to be significant but no longer with the Holm adjustment, Green: C scored lower than cue, Blank: no significant difference in scores.

	All	50-50	75-25	100-0
A1	$t(59) = 1.46$, $p = .45$	$V = 115$, $p = 1$	$V = 52$, $p = 1$	$V = 92$, $p = 1$
B1	$t(59) = 3.09$, $p = .02$	$V = 120$, $p = .54$	$V = 31$, $p = 1$	$V = 140$, $p = .05$
X1	$t(59) = 2.82$, $p = .003$	$V = 175$, $p = .54$	$V = 84$, $p = 1$	$V = 151$, $p = .98$
C	$t(59) = -0.98$, $p = .45$			
A3	$t(59) = 5.18$, $p < .001$	$V = 82$, $p = .20$	$V = 181$, $p = .34$	$V = 200$, $p = .26$
B3	$t(59) = -2.56$, $p = .05$	$V = 9$, $p = .11$	$V = 61$, $p = 1$	$V = 92$, $p = 1$
X3	$t(59) = 1.35$, $p = 0.45$	$V = 38$, $p = 1$	$V = 113$, $p = 1$	$V = 156$, $p = 1$

Table 3.6: Differences scores to fossil and diamond Experiment 1. Red: used to be significant but no longer with the Holm adjustment, Green: fossil is scored lower than diamond, Blank: no significant difference in scores. V scores are for Wilcoxon tests, t scores for t-tests.

Difference in scores to a new cue: In Table 3.5 we can see the results of the Wilcoxon tests comparing the scores given to a cue in the test phase and the scores given to the new cue C in the test phase. A green square indicates that C was scored lower than that cue, a red square indicates that the difference between scores was significant before the Holm p-value correction, and a blank square indicates that there was no significant difference.

We first look at the cues of the first block. For cue A there was no significant difference between the scores, not for the overall score and not for any of the separate conditions either. Cue B was scored higher than cue C when we look at the overall scores, but over conditions there were no significant differences. Cue X scored higher than the new cue overall, and also in condition 50-50 and 100-0, but not in condition 75-25.

For the cues of the last block, cue B was never scored significantly different from the new cue C. Cue A however was always scored higher, regardless of condition. Cue X scored higher than cue C overall and in the 75-25 and 100-0 condition.

There did not seem to be a generalizable pattern, however as most cues scored significantly higher than the new cue C, these results support the Rescorla-Wagner model as well.

Difference in scores to a new outcome: Aside from a new cue, we also introduced a new outcome in the test phase. While we did not model this in our simulations, it is still of interest to look at. This is because if there was a significant difference between the scores to diamond and to fossil, then that would mean that participants did succeed in learning a connection between the cues and the diamond outcome.

The results of the different tests can be seen in Table 3.6, where a t-test was used when the difference between data was normally distributed, and a Wilcoxon test was used when this data was not normally distributed. Green squares indicate that diamond was scored higher than fossil for that cue, red indicates a result that was significant before applying the Holm adjustment to the p-values and the blank squares were not significantly different.

If we look at the overall scores we can see that cues B and X from the first block were scored significantly higher to diamond than to fossil, as was cue A from the third block. However this is no longer the case when looking at the results over conditions. This indicates that maybe participants did not learn a strong connection between diamond and the cues in the training phase, as in all conditions they score the same to a completely new outcome. It is also interesting to see that cue C did not score significantly different to fossil and diamond, which is as we expect, as for cue C these two outcomes are both new outcomes.

3.3 Discussion

This experiment aimed to replicate the findings of Van Hamme and Wasserman, as well as tried to disentangle which of the two model predictions would best describe the behaviour found in the test phase of our simulations.

If in the test phase the cues were scored equally as high as in their original block, this would be in line with the Rescorla-Wagner model. When the cues of the first block would be scored lower in the test phase as in the original blocks, and the cues of the third block scored equal, then this would be in line with the Van Hamme-Wasserman model. We also made predictions of the cues in relation to a newly introduced cue. The Rescorla-Wagner model predicted that the cues of the first and third block would be more predictive of an outcome than a new cue, while the Van Hamme-Wasserman model predicts that cues from the first block would be less predictive of an outcome than a new cue and the cues from the third block would be more predictive of an outcome.

We found support in the direction of the Rescorla-Wagner model, as the scores between the training and test phase did not differ, expect for the overall score of cue A in the first block. The scores to the new cue to a diamond outcome were lower than almost all the cues from the training phase when looking at the over all scores. When looking at the separate conditions there does not seem to be a clear pattern. The over all scores also support the Rescorla-Wagner model, as both the cues from the first and the third block score higher than a new cue.

When comparing the results of our training phase to that of Van Hamme and Wasserman's results we find a similar kind of effect over condition, but a lot less strong. We also found that participants did not seem to know very well how to score, as Figure 3.7 showed that participants still rate a cue that always leads to a diamond outcome, as not leading to a diamond outcome most frequently. The fact that participants might not have know well how to score, is also supported by the fact that participant did not score diamond trials higher than no diamond trials for most of the conditions and cues.

It could be possible that our findings were the result of other external factors. One of those factors is something that could be seen in the test phase, when looking at the connections to diamond and fossil. While some of the overall scores did

indicate that diamond was scored differently from fossil, when split over conditions, diamond was never scored different from fossil. This means that a completely new outcome has the same causal rating score as an outcome that has already been encountered before. This could indicate that there was no relationship learned between these cues and the diamond outcome in the first place.

Previous research by Garcia et al. (1968) has shown that a cue must be appropriate for the outcome, meaning that if a participant (or in their experiment, a rat) did not think that a certain cue could lead to an outcome in the first place, then there would be no connection formed and there would be no weights to update. It could be that participants did not believe that, for example, an item such as a log could ever lead to a diamond, thus resulting in no connection forming between the two in the first place.

In our next experiment we therefore want to make it more salient that these objects do not hold the same properties or relationships as they do in 'our' world.

Another issue that we encountered, was the fact that participants gave a very low causal rating score in general, but most curiously to a cue that always predicted a diamond outcome. This might be one of the reasons as to why our effect of condition was so small compared to that of Van Hamme and Wasserman, as it could have resulted from the fact that participants did not fully understand how to score the experiment. Therefore in the following experiment we would like to give participants more clear instructions and add a survey at the end of the experiment asking them whether these instructions were clear, what they did and did not understand and more.

One more aspect that can be improved in the next experiment is the fact that there might not have been a clear enough visual distinction between the three outcomes. The diamond, fossil and no diamond outcome all featured a drawing of the same dirt hole. The only difference was whether or not a diamond or fossil occurred inside of it. We want to make these three outcomes more visually distinct. As the diamond trials were not always scored higher than the no diamond trials, this could indicate that participants did not distinguish the two different outcomes, Since they did also not seem to notice that cues predicted outcomes in difference probabilities, we hope that increasing clarity on the outcomes will result in more clarity of the relations between the cues and outcomes as well.

To conclude, while we did find evidence for the Rescorla-Wagner model, we did not fully replicate the findings of Van Hamme and Wasserman and the scoring to the experiment also did not follow our expectations. Therefore we will run a second experiment. In this new experiment we will introduce clearer instructions, a more intuitive practice run and a greater distinction between the two outcomes. We hope that these changes will make it more salient to the participants what they will have to do. We will also add a survey at the end of the experiment for exploratory research, to get more insight on if participants found the instructions clear or not.

4 Experiment 1B

We ran Experiment 1 again, however this time we made adjustments to the explanation and training phase, to hopefully make it more explicit to participants what we expect of them.

As the methods of this experiment are largely the same as those described in Section 3.1, we will focus on what we changed in the current experiment as opposed to our Experiment 1.

4.1 Methods

4.1.1 Participants

The participants were recruited via the online platform Prolific. The selection criteria were the same as in the previous experiment, with the addition of excluding participants that took part in the previous experiment.

In total 15 participants took the experiment. Five of which returned their experiment (thus indicating that they no longer wanted to participate). Therefore we had a total of 10 participants. They were paid £2.50 and were told that the experiment would take 20 minutes (the average completion time was 21 minutes).

4.1.2 Materials/Stimuli

We updated the outcomes compared to the previous experiment. These changes can be seen in Figures 4.1 and 4.2. We wanted to make the outcomes more distinct by having the diamond be a picture on its own, instead of it appearing in the hole. The same was done for the fossil outcome, while the no diamond outcome is still just an empty hole.

4.1.3 Experimental Design

While the experimental design was largely the same compared to Experiment 1, we did change a few things to improve the clarity of the explanation.

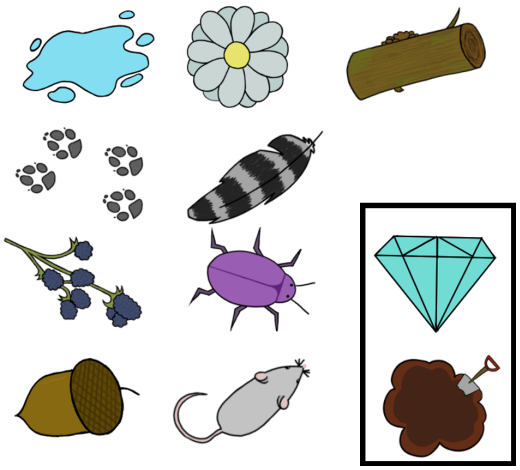

Figure 4.1: Cues and outcomes shown in the training phase of Experiment 1B. Outcomes are in the black box.

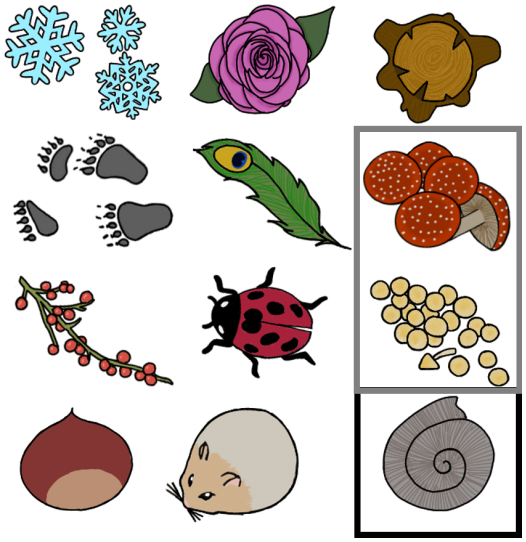

Figure 4.2: Cues and outcomes shown in the test phase of Experiment 1B. Outcomes are in the black box.

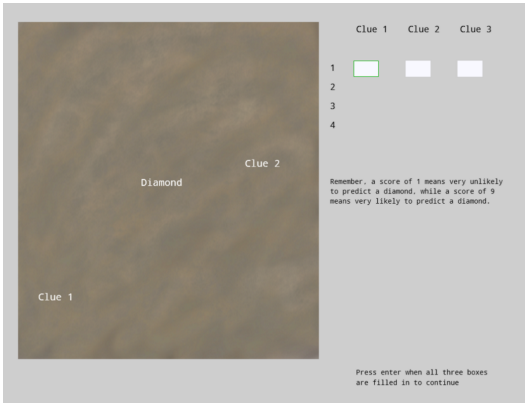

Figure 4.3: Example of a screen that a participant saw in practice of Experiment 1B.

First, we changed the explanation text to include less backstory about being a space adventurer. We wanted to get straight to the point and to avoid participants getting confused by irrelevant details. We also put more focus on telling them that the connections between objects they already knew, might not exist on this planet. This was done to make it clear that connections between these objects and a treasure was possible.

The biggest change was made to the practice phase. As can be seen in Figure 4.3, we added 'cues' that appear on screen. This means that not only could participants practice rating, they can also get used to the fact that the cues will change location, and that the outcome might change per trial.

We have also added small blocks of text in the practice phase that remind people that the outcome could also be *a lack* of a diamond, and reminded them what the rating scale was.

Lastly we introduced a survey at the end of the test phase. This survey was made in Qualtrics (https://www.qualtrics.com). It contained six questions all related to whether participants understood the experiment, what they rated according to, if their rating strategy changed and more. The questions can be found in Appendix B. The survey was added for exploratory reasons and the exact answers will not be discussed further, however we will present whether or not participants understood the instructions of the experiment in the results.

4.1.4 Procedure

The procedure for this experiment was the same as Experiment 1. The only difference is that once the experiment ended, they were automatically redirected to Qualtrics, where they could answer the questions to the survey. On average people spend 2 and a half minutes on the survey.

4.2 Results

The results are split by training phase and test phase. In the training phase we compare the results of the current experiment to those of Van Hamme and Wasserman. The second section concerns the test phase, where we can check our model predictions. We also added a third section, where we will discuss the results of the survey. In this results section we will only discuss results that answer our main research question, and results that indicate if we improved upon our explanation (one of the goals of this experiment). Analysis that are done outside of these restrictions are discussed in Appendix A.

It is important to note that the group of participants is much smaller in this experiment (10 participants) when compared to the Van Hamme and Wasserman experiment (48) or our Experiment 1 (60).

4.2.1 Training Phase

We will first compare the results of Van Hamme and Wasserman's paper to our own results. Figure 4.4 contains multiple plots, comparing our results to that of Van Hamme and Wasserman. For all of these plots, the data from Van Hamme and Wasserman is on the left, and that of the current experiment is on the right.

We will first discuss the average causal rating scores per condition, which is shown in the plot in the top left of Figure 4.4. In these plots the x-axis displays the three conditions (50-50, 75-25 and 100-0) and on the y-axis the causal rating score can be seen. Note that this causal rating score ranges from 0 to 8 in the original experiment, but in our current experiment participants rated from 1 to 9.

On the left we see that the higher the chance that cue A leads to an allergic reaction (in our experiment a diamond), the higher the average score is, while cues B and X decrease in score. In our experiment, while cue A shows an increase and B and X a slight decrease, these changes are less pronounced as in the Van Hamme and Wasserman plot. Compared to our previous experiment the three cues are scored more varying scores, however as cue X is scored higher than both other cues it could still indicate that participants did not understand what they had to do in this experiment.

It is also of interest to look at the individual conditions and investigate what happens over trials, as we still might see a similar pattern over trials emerge as in the Van Hamme and Wasserman results.

In the top right of Figure 4.4 we see the average causal rating score for all participants over trials for condition 50-50. The grey line in the left plot separates the pre-scoring, from the rest of the trial data.

While we see a similar pattern between the two experiments, Experiment 1B has a lot more fluctuation within the scores of all the cues. This could be due to the fact that there are a lot less participants.

The bottom left of Figure 4.4 shows the average causal rating score for all participants over trials for condition 75-25. In the left plot we can see participants started to learn that over trials that cue B is a worse predictor of an allergic reaction than the other two cues, while A is a better one. In the right plot we do see that B is scored lower overall than the two other cues, and that A increases over trial, but cues A and X are still very much entangled. Once again there is a lot more fluctuation in scores in our experiment.

Lastly the bottom right part of Figure 4.4 shows the average causal rating score for all participants over trials for condition 100-0. In the left plot we now see a very clear distinction between cues. Cue A is scored very high, as it always predicts an allergic reaction, while B is learned to be not very predictive (or very predictive of a lack of an allergic reaction) and thus gets a lower score. As X is shown with both cues A and B, this cue's score also drops, as it is not as predictive of an allergic reaction as cue A is. In our experiment data on the right, we do see a slightly lower score for B when compared to the other two cues. However this distinction is a lot less clear as in the results of Van Hamme and Wasserman. We also do not get the distinction between cues A and X that the original paper found.

In Figure 4.5 the frequencies of the causal rating scores for cues A, B and X are shown. In the top left we see condition 50-50, in the top right condition 75-25 and on the bottom condition 100-0. We will not discuss all of the plots in this Figure, but we will highlight those that are of interest.

In condition 50-50 both AX and BX lead to a diamond in half of their trials, and to no diamond in the other half. One would then expect that the scores given to cues A and B are either quite varied, or that they are given as score of 5 (possibly) most often, as it is quite unsure if they led to a diamond or not. However, as also could be seen in Experiment 1, in the top left of Figure 4.5 we can see that both cues A and B were scored a 1 (definitely not leading to a diamond) most often. In condition 100-0, where AX always led to a diamond outcome, while BX always led to a no diamond outcome, one would expect A to be scored a score of 9 (definitely leads to a diamond) most often and B to be scored a 1 most often. While B is scored in line with what we would expect, we see that cue A is still scored a

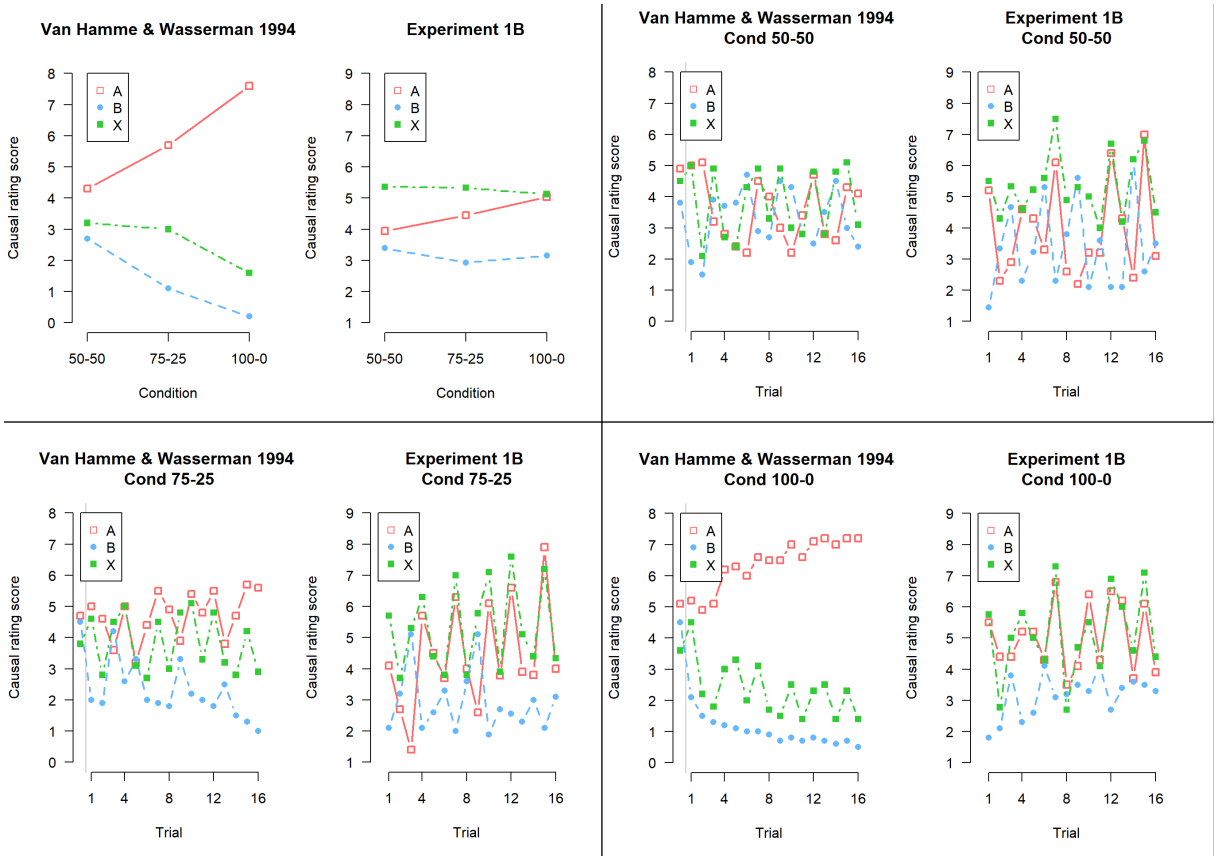

Figure 4.4: Average scores of cues A, B and X of Van Hamme and Wasserman's 1994 paper (left) and that of Experiment 1B (right). Top left: average causal rating scores over conditions, top right: average causal rating scores over trials for condition 50-50, bottom left: average causal rating scores over trials for condition 75-25, bottom right: average causal rating scores over trials for condition 100-0. The grey line vertical line in the plots over trials separates the scores for the pre-scoring from the rest of the trials.

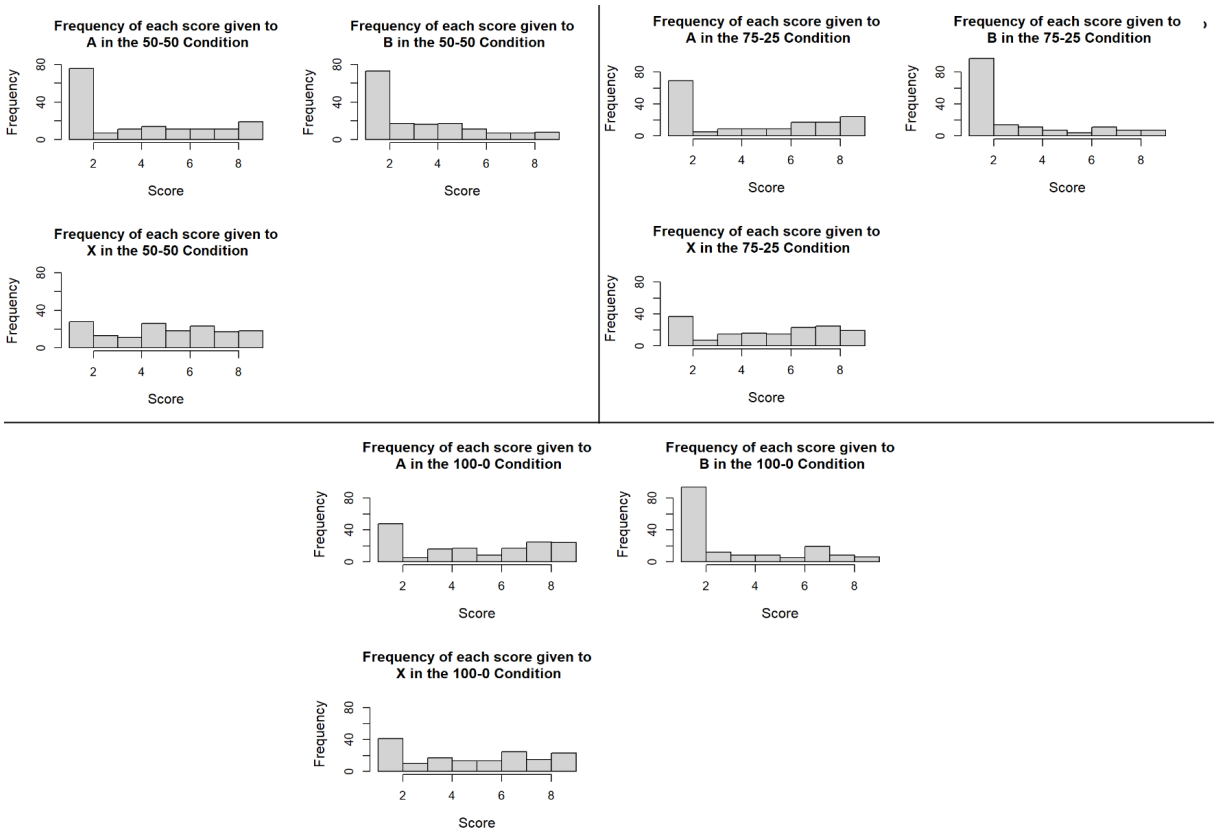

Figure 4.5: Frequency of each of the scores given to cues A, B and X of Experiment 1B. Top left: condition 50-50, top right: condition 75-25, bottom: condition 100-0.

1 most frequently. Compared to Experiment 1, it is scored a score of 8 and 9 more often, but 1 is still the most frequently given score to a cue that always predicts a diamond. It seems then that participants still did not fully understand the scoring.

As the scores differ less per condition as those of Van Hamme and Wasserman, and the participants did not seem to fully understand the scoring, we will now look at whether or not participants scored a diamond trial higher than a no diamond trial. If there was a difference between these two scores, it would suggest that participants at least understood how they had to score on a trial-to-trial basis.

We compared the scores on the no diamond trials to those on the diamond trials with paired Wilcoxon rank tests (as the difference in data was not normally distributed). Because we performed multiple Wilcoxon tests over the same data-set, we adjusted the p-values according to the Bonferroni-Holm method to reduce the probability of obtaining Type I errors. These adjusted p-values are also the ones that we will report. The averages of the scores for the diamond and no diamond trials can be seen in Figure 4.6. This Figure depicts the scores for the cues in all the trials, so trials in which the cue appeared, but also those in which the cue was

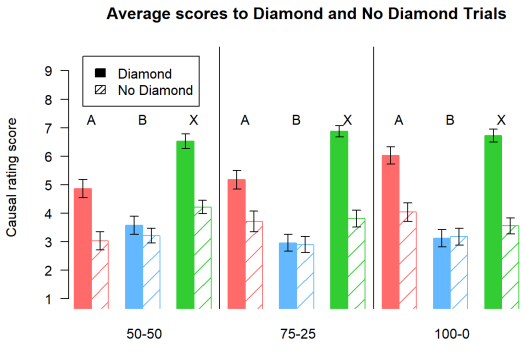

Figure 4.6: The average scores to each condition for each cue on diamond (solid colour) trials and no diamond (hatched) trials for Experiment 1B. The error bars represent the standard error.

absent.

For cue A diamond was scored higher than no diamond trials in condition 100-0 ($V = 55$, $p = .02$), but there was no significant difference in condition 50-50 ($V = 52$, $p = .07$) or 75-25 ($V = 49$, $p = .11$). Cue B was never scored significantly different, not in condition 50-50 ($V = 29$, $p = 1$), 75-25 ($V = 20$, $p = 1$), nor 100-0 ($V = 16$, $p = 1$). For cue X diamond was scored higher than no diamond trials in condition 50-50 ($V = 55$, $p = .02$), but there was no significant difference for conditions 75-25 ($V = 52$, $p = .06$) and 100-0 ($V = 45$, $p = .06$). So overall only cue A in condition 100-0 and cue X in condition 50-50 scored higher on a diamond trial than a no diamond trial. Therefore is seems as though, on a trial to trial basis, participants might not have understood what exactly they had to score to.

4.2.2 Test Phase

The test phase was introduced to test the predictions obtained through our computational simulations, which we will repeat here for clarity before discussing the results.

In Figure 2.3 we saw that for the weights to allergic, the Rescorla-Wagner model predicted that cues that were asked in the first block still had the same connection weight when they were asked in our modelled test phase. The same can be seen for the connection weights of the cues from the last block in the training phase that participants saw. We assume here that this connection weight is related to the scores, so that the scores will also stay the same.

In Figure 2.3 we could also see that the Van Hamme-Wasserman model predicts a decrease in connection weights to allergic for cues from the first block when they are seen again in the test phase. This decrease is not seen for cues from the last block, as those have not been able to decrease in weight yet.

We also looked at the connection weights related to a new cue introduced in the test phase. The Rescorla-Wagner model predicts that cues that have been seen before in the training phase, have a higher connection weight to allergic than a cue that has not been seen before. In the Van Hamme-Wasserman model, the cues from the first block have a lower connection weight than the new test phase cue, but cues from the third block are scored higher.

We will check which of these predictions holds by looking at the first two parts of our test phase. As there was no difference found between the scores in these two parts, we combined them into one data set. As we perform multiple t-tests over the same data-set, we use and report on the adjusted p-values according to the Bonferroni-Holm method for this whole result section.

Due to the fact that the cues behaved differently depending on condition, for example, cue A predicted a diamond more often in 100-0 than in 50-50 and would thus be scored differently depending on the condition, we will look at both the effect of the cues overall and the effect of the cues per condition.

As mentioned in the beginning of the results section, we will only report on the results that help answer our research question and compare our research to that of van Hamme and Wasserman in this section of the thesis. The results of the third part of the test phase and the comparison between part 1 and 2 can be found in Appendix A.

Difference in scores training and test phase: Table 4.1 contains the results of the Wilcoxon tests between the scores to the cues in the training phase and the test phase. We used a Wilcoxon test instead of a t-test, as the difference in data was not normally distributed. The squares that are coloured red are cues that were significant but with the Holm correction are no longer and the blank squares were not significantly different.

None of the cues, regardless of looking at the overall score or to the separate conditions, differ significantly their scores to the test and training phase. This would support the Rescorla-Wagner model, as this model predicted that there would be no significant difference between the scores in the training and in the test phase for both blocks.

Difference in scores to a new cue: In Table 4.2 we can see the results of the Wilcoxon tests comparing the scores given to a cue in the test phase and the scores given to the new cue C in the test phase. A green square indicates that C was scored lower than that cue, a red square indicates that the difference between scores was significant before the Holm p-value correction, and a blank square indicates that there was no significant difference.

We first look at the cues of the first block. For cue A and B there was no significant difference between the scores, not for the overall score, nor for the separate conditions. Cue X was scored higher than cue C when we look at the overall score and all separate conditions except for condition 100-0.

For the cues of the last block, cues B and X were never scored significantly different from the new cue C. Cue A however was is scored higher in the overall score, as well as in condition 100-0.

Overall there did not seem to be a generalizable pattern, however as almost none of the cues score significantly higher than the new cue C, it is difficult to say which of the two models these results support. In our predictions for both models we expected a

	All	50-50	75-25	100-0
A1	$V = 5$, $p = .19$	$V = 3$, $p = 1$	$V = 0$, $p = 1$	$V = 3$, $p = 1$
B1	$V = 36$, $p = .55$	$V = 5$, $p = 1$	$V = 10$, $p = 1$	$V = 2$, $p = 1$
X1	$V = 50$, $p = .12$	$V = 7$, $p = 1$	$V = 10$, $p = 1$	$V = 3$, $p = 1$
A3	$V = 12$, $p = .70$	$V = 1$, $p = 1$	$V = 2$, $p = 1$	$V = 4$, $p = 1$
B3	$V = 20$, $p = 1$	$V = 0$, $p = 1$	$V = 6$, $p = 1$	$V = 1$, $p = 1$
X3	$V = 27$, $p = 1$	$V = 3$, $p = 1$	$V = 6$, $p = 1$	$V = 5$, $p = 1$

Table 4.1: Differences in scores to cues in the training phase and test phase of Experiment 1B. Red: used to be significant but no longer with the Holm adjustment, Blank: no significant difference in scores.

	All	50-50	75-25	100-0
A1	$V = 4$, $p = .10$	$V = 15$, $p = .38$	$V = 17$, $p = .38$	$V = 8$, $p = .47$
B1	$V = 4$, $p = .10$	$V = 13$, $p = .27$	$V = 12$, $p = .27$	$V = 4$, $p = .38$
X1	$V = 0$, $p = .01$	$V = 2$, $p = .04$	$V = 0$, $p = .02$	$V = 3$, $p = 0.38$
A3	$V = 1$, $p = .04$	$V = 4$, $p = .38$	$V = 5$, $p = .07$	$V = 0$, $p = .02$
B3	$V = 6$, $p = .18$	$V = 21$, $p = 1$	$V = 18$, $p = .38$	$V = 29$, $p = .77$
X3	$V = 2$, $p = .06$	$V = 1$, $p = .24$	$V = 11$, $p = .23$	$V = 5$, $p = .07$

Table 4.2: Differences in scores to cues in the test phase and scores to the new cue C for Experiment 1B. Red: used to be significant but no longer with the Holm adjustment, Green: C scored lower than cue, Blank: no significant difference in scores.

	All	50-50	75-25	100-0
A1	$V = 37$, $p = .29$	$V = 6$, $p = 1$	$V = 4$, $p = 1$	$V = 3$, $p = 1$
B1	$V = 32$, $p = .23$	$V = 5$, $p = 1$	$V = 5$, $p = 1$	$V = 3$, $p = 1$
X1	$V = 55$, $p = .01$	$V = 10$, $p = 1$	$V = 10$, $p = 1$	$V = 3$, $p = 1$
C	$V = 4$, $p = .41$			
A3	$V = 54$, $p = .05$	$V = 3$, $p = 1$	$V = 9$, $p = 1$	$V = 10$, $p = 1$
B3	$V = 23$, $p = 1$	$V = 1$, $p = 1$	$V = 9$, $p = 1$	$V = 2$, $p = 1$
X3	$V = 48$, $p = .19$	$V = 3$, $p = 1$	$V = 8$, $p = 1$	$V = 7$, $p = 1$

Table 4.3: Differences scores to fossil and diamond Experiment 1B. Red: used to be significant but with Holm not anymore, Green: fossil is scored lower than diamond, Blank: no significant difference in scores.

difference, the direction of that difference changed according to the specific model.

Difference in scores to a new outcome: Aside from a new cue, we also introduced a new outcome in the test phase. While we did not model this in our simulations, it is still of interest to look at. This is because if there was a significant difference between the scores to diamond and to fossil, then that would mean that participants did succeed in learning a connection between the cues and the diamond outcome.

The results of the different tests can be seen in Table 4.3, where a Wilcoxon test was used as this data was not normally distributed. Green squares indicate that diamond was scored higher than fossil for that cue, red indicates a result that was significant before applying the Holm adjustment to the p-values and the blank squares were not significantly different.

If we look at the overall scores, we can see that only cue X from the first block is scored significantly higher to diamond than to fossil. However this is no longer the case when looking at the results over conditions. This indicates that participants did not learn a strong connection between diamond and the cues in the training phase, as in all conditions they score the same to a completely new outcome. It is also interesting to see that cue C did not score significantly different to fossil and diamond, this is in line with what we expect, as for cue C these two outcomes are both new.

4.2.3 Survey Results

The first question of the survey participants answered was "Did you find the instructions at the beginning of the experiment clear?". Participants could answer this open question with any amount of characters. The goal of this question was to see if we did indeed improve the instructions of our experiment with the adjustments we made.

Less than half of all participants (46%) indicated that they understood the instructions, 23% was unsure if they understood them and more than a quarter of all participants (31%) indicated that they did not understand at all. This seems to indicate that, although we made certain aspects of our explanation more clear, most participants still did not understand what they had to do in the experiment.

4.3 Discussion

This experiment aimed to replicate the findings of Van Hamme and Wasserman, as well as trying to disentangle which of the two model predictions would best describe the behaviour found in the test phase of our simulations. In addition to our first experiment, we also aim to check if the instructions of our experiment were clear.

If in the test phase the cues were scored equally as high as in their original block, this would be in line with the Rescorla-Wagner model. When the cues of the first block would be scored lower in the test phase as in the original blocks, and the cues of the third block scored equal, then this would be in line with the Van Hamme-Wasserman model. We also made predictions of the cues in relation to a newly introduced cue. The Rescorla-Wagner model predicted that the cues of the first and third block would be more predictive of an outcome than a new cue, while the Van Hamme-Wasserman model predicts that cues from the first block would be less predictive of an outcome than a new cue and the cues from the third block would be more predictive of an outcome.

We found partial support in the direction of the Rescorla-Wagner model, as the scores between the training and test phase did not differ. The scores to the new cue to a diamond outcome however, did not differ except for a few cues. Therefore the scores to the new cues do not seem to support the Rescorla-Wagner model, nor the Van Hamme-Wasserman model, as for both predictions we need a difference in scores.

When comparing the results of our training phase to that of Van Hamme and Wasserman's results, we find a similar kind of effect over condition, but a lot less strong. Cue X is also scored higher than the most predictive cue A (Figure 4.4, top left), which in the original experiment was the other way around. We also found that participants still did not seem to know very well how to score, as Figure 4.5 showed that participants still rate a cue that always leads to a diamond outcome, as not leading to a diamond outcome most frequently. This is also supported by the fact that participant did not score diamond trials higher than no diamond trials for most of the conditions and cues.

One of our goals of this experiment was to increase the clarity of the explanations. Less than half of the participants indicated that they had clearly understood what they had to do in the experiment. More than a quarter even indicated that they did not understand the instructions at all. This indicates that we did not yet succeed in increasing the clarity of the experiment.

It could be possible that our findings were the result of other external factors. One of those factors is something that we also found in our Experiment 1. When looking at the results from the test phase, we saw that diamond was almost never scored higher than fossil. This could indicate that there was no relationship learned between these cues and the diamond outcome. Combining this with the results

of our survey, this indicates that the combination of cues and outcomes might still be unclear for participants.

We also still found that participants gave a very low causal rating score in general, most curiously to the cue that always predicted a diamond outcome.

To conclude, while we did find partial evidence for the Rescorla-Wagner model, we did not fully replicate the findings of Van Hamme and Wasserman and the scoring to the experiment still did not follow our expectations. The survey results, paired with the small sample group calls for more research. Therefore we will now perform a replication of the original experiment of Van Hamme and Wasserman's 1994 paper. As we found that there seems to be no connection made between the diamond outcome and the cues in both this experiment and the previous one, we will use the food paradigm for the next experiment, since we know for sure that there is a connection between foods and allergic reactions.

5 Experiment 2

As could be seen in the previous two experiments, we have not been able to replicate the findings of the original 1994 paper with our adjustments in place. Therefore we want to create a full replication of the original experiment, the only difference being that the current experiment will be run online.

5.1 Methods

5.1.1 Participants

The participants were selected via the online platform Prolific, we adhered to the same selection criteria as in Experiment 1B.

In total 32 participants took part in the experiment, of which 10 participants returned their submission (and thus indicated that they no longer wanted to participate) and two timed out. This left 20 participants, of which one did not fully finish the experiment, as it crashed. This participant was still included in the analysis for the parts that they did complete.

The participants received a monetary reward of £2.50 and they were told it would take 20 minutes (average completion time of 26 minutes).

5.1.2 Materials/Stimuli

The stimuli were kept the same as in the experiment of Van Hamme and Wasserman (1994). This means that instead of pictures, as in our previous experiments, they would now just see food in a written form. Which foods they would see, was dependent on which group a participant was sorted into. Each participant saw three out of the six food groups (Table 5.1) in the training phase.

Food group	Condition	A	B	X
1	50-50	Strawberries	Peanuts	Shrimp
2	75-25	Bran	Cabbage	Yogurt
3	100-0	Chicken	Mustard	Bananas
4	50-50	Walnuts	Peaches	Wheat
5	75-25	Horseradish	Lobster	Corn
6	100-0	Cheese	Pork	Blueberries

Table 5.1: The different food groups, their outcome condition and which foods are filled in for cue A, B and X.

Food	Alternative	Food	Alternative
Strawberries	Blueberries	Cabbage	Corn
Chicken	Pork	Peaches	Bananas
Horseradish	Mustard	Lobster	Shrimp
Bran	Wheat	Cheese	Yogurt
Walnuts	Peanuts	Fennel	Celery

Table 5.2: The cues and their alternatives. Paired by similarity, food groups 1, 2 and 3 were paired with 4, 5 and 6 and the other way around. Fennel is the new cue C and celery is the alternative, both of these were not in the original food groups.

As for the test phase, participants saw alternative cues based on the ones they had seen in the training phase. These alternative cues were taken from the food groups that a participant did not see in the training phase, which were matched to the ones that they did see. Which foods were matched with which alternatives can be found in Table 5.2. The only fully new cues were those used for the new cue C (fennel) and its alternative (celery).

5.1.3 Experimental Design

There were a total of six food groups. We tried to stay as close as possible to the way that these food groups were divided over participants compared to the Van Hamme and Wasserman experiment. Instead of randomising stimuli over participants, as we did in the previous two experiments, we divide participants into six groups, depending on their subject number. The order of food conditions depended on which of these groups a participant belongs to, and is based on Table 2 from the Van Hamme and Wasserman paper (see Table B.1 in Appendix B for a replication of the table).

Participants had to rate the likelihood of the cues leading to an allergic reaction on a scale of 0 to 8. This time both the number pad (numpad) and the keyboard could be used. This scale is now explicitly shown in the explanation (see for an example Figure 5.1), with indications showing what a 0 (definitely not), a 4 (probably) and an 8

(definitely) mean.

The training phase still consisted of three times 16 trials. However, before each of the 16 trials, the participants were asked how likely they thought each of the three foods that they would see in the next block, would lead to an allergic reaction. This is the same pre-scoring method that Van Hamme and Wasserman used.

The timing also differed compared to our previous two experiments. Instead of participants seeing the cues first, then the outcome with the cues and only then can they also rate, now everything appears at once. This means that participants see the cues, outcomes and rating system for the full 15 seconds of a trial.

In addition to the three parts of the test phase from the previous experiments, we added a fourth part at the end. In this part of the test phase, participants saw one food on the screen and had to rate for either an allergic reaction or a fever (the new outcome in the test phase), how likely they think it is that they will be the result of this food. This part was added to see if there would be a big difference between the pre-scores given to the cues of the first and the last block and the rating to the allergic reaction in the test phase. This part of the test phase consisted of seven cues (three from the first block of the training phase, three from the last and one novel cue).

Participants saw a total of four trials in the practice phase, three blocks of 16 trials plus three pre-score trials (51) in the training phase, and 42 trials in the test phase. This means that they saw a total of 97 trials.

After these trials, participants received the same survey given in Experiment 1B. The average completion time on the survey was 4 minutes.

5.1.4 Procedure

The procedure was similar to that of Experiment 1 and 1B with some small changes. Those changes will be discussed here.

Once participants chose to take part in the experiment and agreed to the consent form, they were given an explanation of the experiment. In short, they were allergists and their patient would eat three foods over the course of the next 16 days. Participants would see which two foods a patient ate on any given day and if there was an allergic reaction that day or not. Every 16 days, a patient would eat something different and each "day" corresponded to one trial. Participants were not told about the testing phase.

Participants could practice with the rating system after the first explanation. This screen looked similar to that of Experiment 1B, except now with the text "Food 1", "Food 2" and "Allergic reaction"

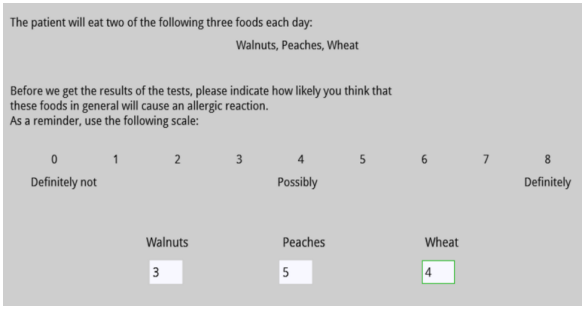

Figure 5.1: A zoomed-in example of a pre-score screen that a participant saw before a training block in Experiment 2.

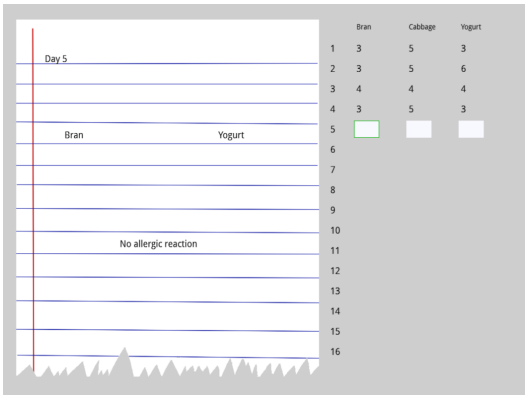

Figure 5.2: Example of a screen that a participant saw in the training phase in Experiment 2.

or "No allergic reaction". The same way of moving between boxes and typing was adopted, with the only difference being that participants could now also rate with the numpad.

Once the practice round was finished participants would first move to the training phase. Instead of starting with the results of the patient right away, they would see a pre-screen (Figure 5.1) in which they were told which foods the patient would eat in the coming days. Participants also had to give an initial rating to those foods, indicating how likely they thought it was that these foods would lead to an allergic reaction in the first place.

After this, participants would continue with the training phase (Figure 5.2) as they would in the previous experiments, with the only difference being that they now saw the cues, outcomes and rating system together for 15000 ms. When participants finished a block, they would see text on screen that told them the next block would start. After this they would see a new pre-scoring screen and continue to the next block.

When a participant finished the training phase, they were informed that they now had to make predictions for the future about the relationships between foods and outcomes. The test phase went in a similar manner as the previous experiments,

with the addition of one new (fourth) part. Before starting this part of the test phase, participants received an extra explanation telling them that they now had to rate according to outcome instead of to cue as they had done previously. So this meant that they would now score the likelihood of an allergic reaction or fever being caused by the food. They could still rate on the same scale.

When participants finished the test phase, they were automatically taken to the Qualtrics website for the survey. The questions asked in this survey were the same ones as in Experiment 1B.

5.2 Results

The results are split by training and test phase. In the training phase we compare the results of the current experiment to those of Van Hamme and Wasserman. The second section concerns the test phase, where we can see whether our model predictions hold. We also added a third section, where we will discuss the results of the survey. In this results section we will only discuss results that answer our main research question, and results that indicate if we improved upon our explanation (one of the goals of this experiment). Analyses that are done outside of that are discussed in Appendix A.

It is important to note that the group of participants, while bigger than in the last experiment, is still smaller (20 participants) when compared to the Van Hamme and Wasserman experiment (48) or our Experiment 1 (60).

5.2.1 Training Phase

We will first compare the results of Van Hamme and Wasserman's paper to our own results. Figure 5.3 contains multiple plots, comparing our results to that of Van Hamme and Wasserman. For all of these plots, the data from Van Hamme and Wasserman is on the left, and that of the current experiment is on the right.

We will first discuss the average causal rating scores per condition, which is shown in the plot in the top left of Figure 5.3. In these plots the x-axis displays the three conditions (50-50, 75-25 and 100-0) and on the y-axis the causal rating score can be seen.

On the left we see that the higher the chance that cue A leads to an allergic reaction, the higher the average score is, while cues B and X decrease in score. In our experiment we see a pattern that is very similar, but the main difference is that the averages scores are less distinctive when compared to that of Van Hamme and Wasserman. This could be the result of participants scoring less extreme on the scale (and thus resulting in an average of the cues that is less distinctive). For that reason, it is also of interest to look at the individual conditions and investigate what happens over trials.

In the top right of Figure 5.3 we see the average causal rating score for all participants over trials for condition 50-50. The grey line in the left plot separates the pre-scoring, from the rest of the trial data. We see a similar pattern between the two experiments, even in terms of fluctuation. All three cues get around the same scores, as they all predict an allergic reaction equally well.

The bottom left of Figure 5.3 shows the average causal rating score for all participants over trials for condition 75-25. In the left plot we can see participants started to learn that over trials that cue B is a worse predictor of an allergic reaction than the other two cues, while A is a better one. On the right we do see that participants score cue B lower than the other two cues and A slightly higher. However cues A and X are still entangled and not as separated in scores as they were in the original experiment.

Lastly the bottom right part of Figure 5.3 shows the average causal rating score for all participants over trials for condition 100-0. In the left plot we now see a very clear distinction between cues. Cue A is scored very high, as it always predicts an allergic reaction, while B is learned to be not very predictive (or very predictive of a lack of an allergic reaction) and thus gets a lower score. As X is shown with both cues A and B, this cue's score also drops, as it is not as predictive of an allergic reaction as cue A is. In our experiment we find a similar pattern, cue A is found to be more predictive of an allergic reaction, B most predictive of a lack of an allergic reaction and cue X scores lower than cue A but still mostly higher than cue B. The difference between the results of the two experiments is that our experiment scores lower on average. Cue A in the left plot eventually reaches an average score of 7, while ours lies more around 6 in the last points. The scores also fluctuate more between trials, but this could be due to the fact that we had a small sample size. So while the pattern is the same, the differences are less extreme, which was also reflected in the average plot in the top left of Figure 5.3.

In the previous two experiments we saw that the frequencies of the scores given to the cues did not match our predictions, based on how often the cues would predict a certain outcome. While the results we found already show a more similar in pattern to that of Van Hamme and Wasserman when compared to the last two experiments, it is still interesting to look at the frequencies of the scores given to each of the cues per condition.

In Figure 5.4 the frequencies of the causal rating

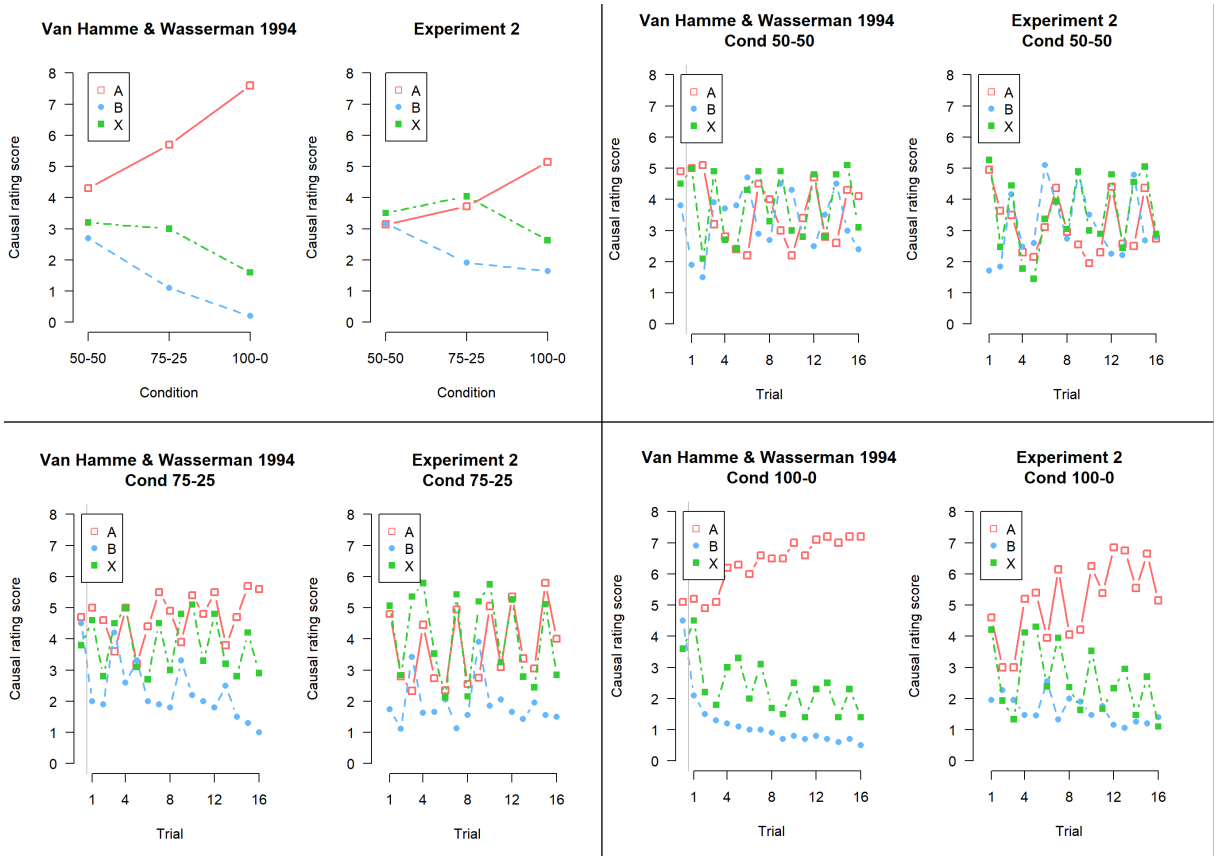

Figure 5.3: Average scores of cues A, B and X of Van Hamme and Wasserman's 1994 paper (left) and that of Experiment 2 (right). Top left: average causal rating scores over conditions, top right: average causal rating scores over trials for condition 50-50, bottom left: average causal rating scores over trials for condition 75-25, bottom right: average causal rating scores over trials for condition 100-0. The grey line vertical line in the plots over trials separates the scores for the pre-scoring from the rest of the trials.

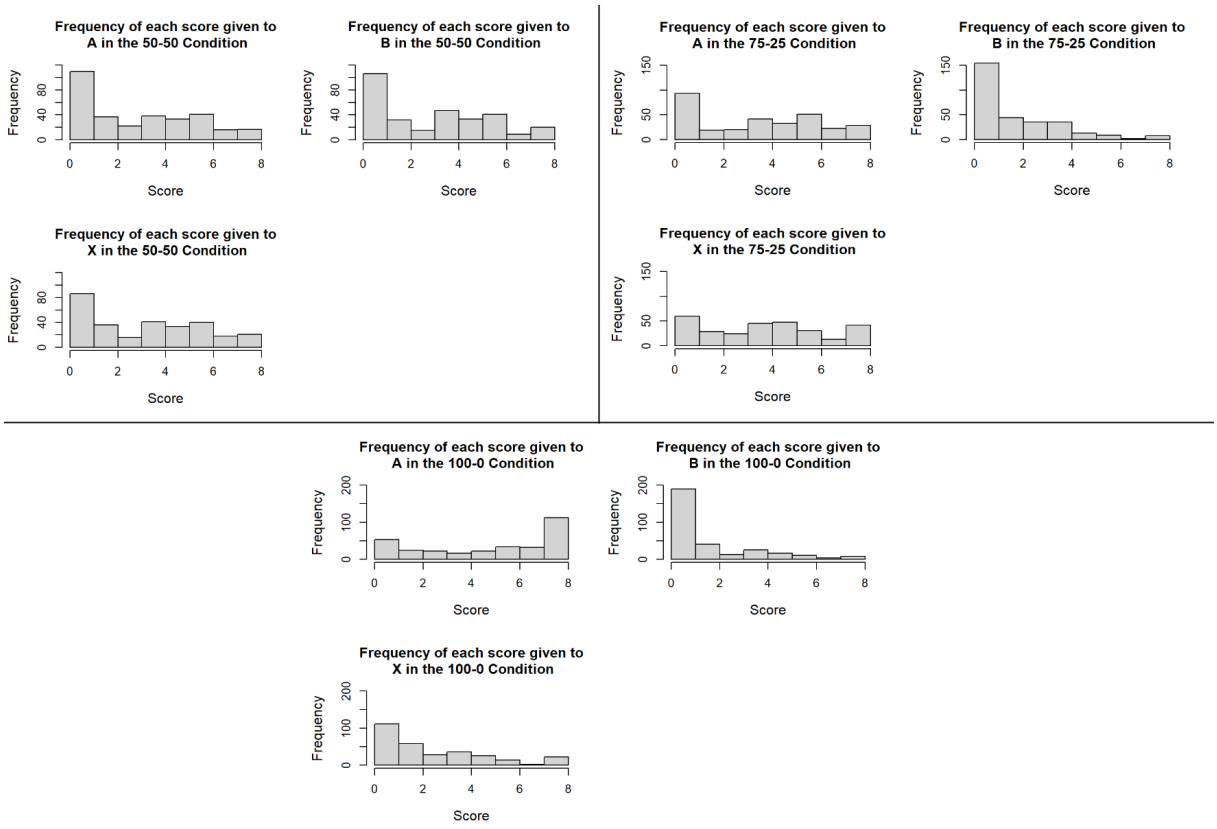

Figure 5.4: Frequency of each of the scores given to cues A, B and X of Experiment 2. Top left: condition 50-50, top right: condition 75-25, bottom: condition 100-0.

scores for cues A, B and X are shown. In the top left we see condition 50-50, in the top right condition 75-25 and on the bottom condition 100-0. We will not discuss all of the plots in this Figure, but we will highlight those that are of interest.

First we will discuss condition 100-0, in which AX always led to an allergic reaction, while BX always led to a lack of an allergic reaction. When looking at the bottom of Figure 5.4, we can now also see this reflected in the frequencies of the scores given in this condition. The most frequent score given for cue B is a score of 0 (definitely not leading to an allergic reaction), while the most frequent score for cue A was a score of 8 (definitely leading to an allergic reaction). This is in line with what we would expect as cue A did indeed definitely lead to an allergic reaction, and cue B never led to one (or always led to a lack of an allergic reaction). However, when looking at condition 50-50 (top left of Figure 5.4), where both AX and BX led to an allergic reaction in 50% of their trials, we would expect the scores to either be quite varied, or to be around a score of 4 (possibly leading to an allergic reaction). While the causal rating scores of 3, 4 and 5 are given more frequently than most other scores for both cues A and B in this condition, a score of 0 is still the most frequent one. The same can be seen in condition 75-25 (top right of Figure 5.4), where one would expect that cue A is scored higher causal rating scores more frequently, as AX predicts an allergic reaction in 75% of all AX trials in this condition. Here we still see however, that a causal rating score of 0 is the most frequent score.

In our previous two experiment participants did not seem to score diamond trials (in this experiment allergic trials) higher than no diamond (not allergic) trials. As our results came closer to those of Van Hamme and Wasserman will investigate if participants also scored the allergic trials higher than the no allergic trials. If there was a difference between these two scores, it would suggest that participants understood what they had to do on a trial-to-trial basis.

We compared the scores on the allergic trials to those on the not allergic trials with two-sided, paired t-tests. As we did multiple t-tests over the same data-set, we adjusted the p-values according to the Bonferroni-Holm method to reduce the probability of obtaining Type I errors. These adjusted p-values are also the ones that we will report. The averages of the scores for the allergic and not allergic trials can be seen in Figure 5.5. This Figure depicts the scores for the cues in all the trials, so trials in which the cue appeared, but also those in

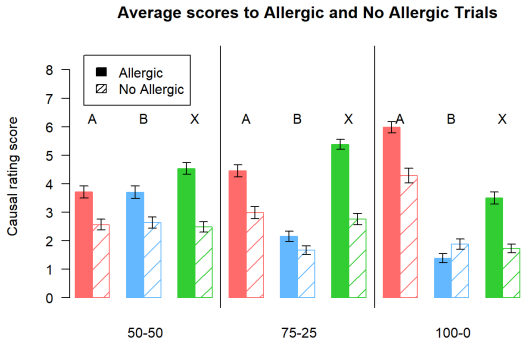

Figure 5.5: The average scores to each condition for each cue on allergic (solid colour) trials and not allergic (hatched) trials for Experiment 2. The error bars represent the standard error.

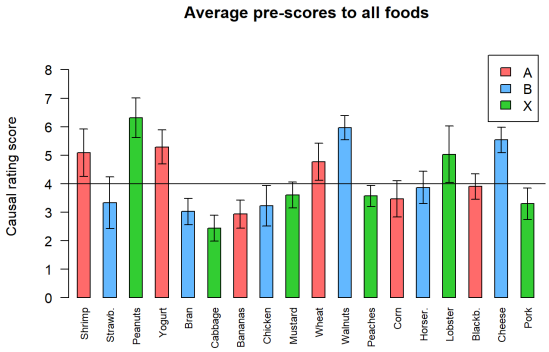

Figure 5.6: The average scores to each food in the pre-scores of Experiment 2. The type of cue of the food is indicated by it's colour. The error bars represent the standard error.

which the cue was absent.

In condition 50-50 cues A ($t(19) = 3.22$, $p = .02$), B ($t(19) = 3.57$, $p = .01$) and X ($t(19) = 3.67$, $p = .01$) were all scored significantly higher on allergic trials than on not allergic trials. In condition 75-25 and 100-0 cues A ($t(19) = 3.61$, $p = .01$; $t(19) = 2.97$, $p = .02$) and X ($t(19) = 4.29$, $p = .004$; $t(19) = 3.73$, $p = .01$) were scored higher on allergic trials than on not allergic trials. There was no significant difference in scores for cue B in condition 75-25 ($t(19) = 2.10$, $p = .10$) and condition 100-0 ($t(19) = -1.47$, $p = .16$).

So except for cue B in conditions 75-25 and 100-0, all cues scored higher on allergic trials than on the no allergic trials. This indicates that participants did understand what they had to score to on a trial-to-trial basis, unlike in the other two experiments. The fact that we did not find this difference for cue B in these two conditions could be because cue B was already less predictive of an allergic reaction in these two conditions. It is then logical that B is scored lower on allergic trials, thus resulting in the scores between allergic and not allergic not being different.

Lastly we looked at the averages given to each food in the pre-scoring, to see if there were any food that were highly indicative of an allergic reaction before seeing results. Van Hamme and Wasserman did not report on any of the pre-scores in their paper, however we still do think it is of interest to investigate if there were already formed connections between allergic reactions and certain foods. In Figure 5.6 we see the average scores given to each of the foods in the pre-score phase. The type of cue a food was is indicated with the colours of the bar graph, red being a cue A, blue a cue B and green a cue X. There is a back line drawn at the causal rating score of 4, which would be a neutral "possibly" score.

We can see that for none of the cue *types* there was a clear high or low score. There are some individual foods that clearly indicate that there is a relation between that food and an allergic reaction, these are shrimp, peanuts, yogurt, wheat, walnuts, lobster and cheese. These are in line with the common allergens that the NHS (NHS, 2019a) describes. There are however also foods that indicate a lack of an allergic reaction, such as bran, cabbage, bananas, and pork.

5.2.2 Test Phase

The test phase was introduced to test the predictions obtained through our computational simulations, which we will repeat here for clarity before discussing the results.

In Figure 2.3 we saw that for the weights to allergic, the Rescorla-Wagner model predicted that cues that were asked in the first block still had the same connection weight when they were asked in our modelled test phase. The same can be seen for the connection weights of the cues from the last block in the training phase that participants saw. We assume here that this connection weight is related to the scores, so that the scores will also stay the same.

In Figure 2.3 we could also see that the Van Hamme-Wasserman model predicts a decrease in connection weights to allergic for cues from the first block when they are seen again in the test phase. This decrease is not seen for cues from the last block, as those have not been able to decrease in weight yet.

We also looked at the connection weights related to a new cue introduced in the test phase. The Rescorla-Wagner model predicts that cues that have been seen before in the training phase, have

a higher connection weight to allergic than a cue that has not been seen before. In the Van Hamme-Wasserman model, the cues from the first block have a lower connection weight than the new test phase cue, but cues from the third block are scored higher.

We will check which of these predictions holds by looking at the first two parts of our test phase. As there was no difference found between the scores in these two parts, we combined them into one data set. As we perform multiple t-tests over the same data-set, we use and report on the adjusted p-values according to the Bonferroni-Holm method for this whole result section.

Due to the fact that the cues behaved differently depending on condition, for example, cue A predicted an allergic reaction more often in 100-0 than in 50-50 and would thus be scored differently depending on the condition, we will look at both the effect of the cues overall and the effect of the cues per condition.

As mentioned in the beginning of the results section, we will only report on the results that help answer our research question and compare our research to that of van Hamme and Wasserman in this section of the thesis. The results of the third part of the test phase and the comparison between part 1 and 2 can be found in Appendix A.

Difference in scores training and test phase: Table 5.3 contains the results of the Wilcoxon tests between the scores to the cues in the training phase and the test phase. We used a Wilcoxon test instead of a t-test, as the difference in data was not normally distributed. The squares that are coloured red are cues that were significant but with the Holm correction are no longer and the blank squares were not significantly different.

If we look at the conditions overall (in the column All), cue B and X seen in the first block were scored lower in the test phase than in the training phase. However when looking at the conditions individually there was no significant difference between the scores in the test and training phase for any of the cues form the first or the third block. While a difference between test and training in the first block, and none in the last block might seem to indicate support for the Van Hamme-Wasserman model, this is not the case. The Van Hamme-Wasserman model predicted a decrease in score instead of an increase as we see here. As there are no significant differences between the scores in the training and test phase when looking at the different conditions, this would go more into the direction of support for the Rescorla-Wagner model, as this model predicted that there would be no significant difference between the scores in the training and the test phase.

Difference in scores to a new cue: In Table 5.4 we can see the results of the Wilcoxon tests comparing the scores given to a cue in the test phase and the scores given to the new cue C in the test phase. A green square indicates that C was scored lower than that cue, a red square indicates that the difference between scores was significant before the Holm p-value correction, and a blank square indicates that there was no significant difference.

We first look at the cues of the first block. For cue A there was no significant difference between the scores when looking at the overall score, however for condition 100-0, cue A was scored higher than cue C. Cue B was scored higher than the new cue overall, but not when looking over conditions. Cue X was scored higher than cue C overall.

For the cues from the last block, cue B was never scored significantly different from the new cue C. Cue A was scored higher for the overall score and condition 50-50, and cue X was not scored differently overall, but was scored higher for condition 75-25.

Overall there did not seem to be a generalizable pattern. However as half of the cues scored significantly higher than the new cue C, and none of them scored significantly lower, these results support the Rescorla-Wagner model.

Difference in scores to a new outcome: Aside from a new cue, we also introduced a new outcome in the test phase. While we did not model this in our simulations, it is still of interest to look at. This is because if there was a significant difference between the scores to allergic and to fever, then that would mean that participants did succeed in learning a connection between the cues and the allergic outcome.

The results of the different tests can be seen in Table 5.5, where a Wilcoxon test was used as this data was not normally distributed. Green squares indicate that allergic was scored higher than fever for that cue, red indicates a result that was significant before applying the Holm adjustment to the p-values and the blank squares were not significantly different.

If we look at the overall scores, we can see that only cues A and X from the third block were scored significantly higher to allergic than to fever. However this is no longer the case when looking at the results over conditions. This indicates that maybe participants did not learn a strong connection between allergic and the cues in the training phase. However, as there already is a pre exisiting connection that we tested and know of, it could also mean that there was not a clear enough distinction

	All	50-50	75-25	100-0
A1	$V = 65$, $p = .74$	$V = 9$, $p = 1$	$V = 3$, $p = 1$	$V = 5$, $p = 1$
B1	$V = 158$, $p = .01$	$V = 10$, $p = 1$	$V = 10$, $p = 1$	$V = 21$, $p = .60$
X1	$V = 144$, $p = .01$	$V = 23$, $p = 1$	$V = 6$, $p = 1$	$V = 21$, $p = .56$
A3	$V = 76$, $p = .74$	$V = 21$, $p = 1$	$V = 4$, $p = 1$	$V = 4$, $p = 1$
B3	$V = 116$, $p = .59$	$V = 6$, $p = 1$	$V = 11$, $p = 1$	$V = 14$, $p = 1$
X3	$V = 140$, $p = .07$	$V = 15$, $p = 1$	$V = 19$, $p = 1$	$V = 20$, $p = 1$

Table 5.3: Differences in scores to cues in the training phase and test phase of Experiment 2. Green: test phase is scored higher than training phase, Red: used to be significant but no longer with the Holm adjustment, Blank: no significant difference in scores.

	All	50-50	75-25	100-0
A1	$V = 23$, $p = .07$	$V = 82$, $p = 1$	$V = 30$, $p = 1$	$V = 4$, $p = .01$
B1	$V = 14$, $p = .047$	$V = 53$, $p = 1$	$V = 17$, $p = .35$	$V = 14$, $p = .08$
X1	$V = 18$, $p = .047$	$V = 43$, $p = .79$	$V = 20$, $p = .50$	$V = 22$, $p = .33$
A3	$V = 13$, $p = .02$	$V = 3$, $p = .004$	$V = 40$, $p = 1$	$V = 25$, $p = .45$
B3	$V = 48$, $p = .78$	$V = 67$, $p = 1$	$V = 64$, $p = 1$	$V = 42$, $p = 1$
X3	$V = 18$, $p = .05$	$V = 40$, $p = 1$	$V = 8$, $p = .03$	$V = 49$, $p = 1$

Table 5.4: Differences in scores to cues in the test phase and scores to the new cue C for Experiment 2. Red: used to be significant but no longer with the Holm adjustment, Green: C scored lower than cue, Blank: no significant difference in scores.

	All	50-50	75-25	100-0
A1	$V = 87$, $p = .16$	$V = 14$, $p = 1$	$V = 5$, $p = 1$	$V = 14$, $p = 1$
B1	$V = 62$, $p = .25$	$V = 6$, $p = 1$	$V = 7$, $p = 1$	$V = 12$, $p = 1$
X1	$V = 98$, $p = .16$	$V = 9$, $p = 1$	$V = 8$, $p = 1$	$V = 21$, $p = .56$
C	$V = 41$, $p = 1$			
A3	$V = 88$, $p = .02$	$V = 21$, $p = .60$	$V = 13$, $p = 1$	$V = 3$, $p = 1$
B3	$V = 39$, $p = 1$	$V = 10$, $p = 1$	$V = 0$, $p = 1$	$V = 8$, $p = 1$
X3	$V = 112$, $p = .04$	$V = 23$, $p = 1$	$V = 20$, $p = .94$	$V = 5$, $p = 1$

Table 5.5: Differences scores to fever and allergic Experiment 2. Red: used to be significant but no longer with the Holm adjustment, Green: fever is scored lower than allergy, Blank: no significant difference in scores.

	Training phase versus Test	Allergy versus Fever	Pre-scoring versus Test
Shrimp	$V = 28$, $p = 1$	$V = 19$, $p = 1$	$V = 20$, $p = 1$
Strawberries	$V = 11$, $p = 1$	$V = 21$, $p = .56$	$V = 3$, $p = 1$
Peanuts	$V = 34$, $p = .60$	$V = 26$, $p = .61$	$V = 13$, $p = 1$
Yogurt	$V = 21$, $p = .60$	$V = 21$, $p = .56$	$V = 15$, $p = 1$
Bran	$V = 9$, $p = 1$	$V = 13$, $p = 1$	$V = 14$, $p = 1$
Cabbage	$V = 11$, $p = 1$	$V = 8$, $p = 1$	$V = 4$, $p = 1$
Bananas	$V = 13$, $p = 1$	$V = 6$, $p = 1$	$V = 0$, $p = 1$
Chicken	$V = 2$, $p = 1$	$V = 3$, $p = 1$	$V = 10$, $p = 1$
Mustard	$V = 7$, $p = 1$	$V = 3$, $p = 1$	$V = 4$, $p = 1$
Wheat	$V = 28$, $p = .28$	$V = 28$, $p = .40$	$V = 13$, $p = 1$
Walnuts	$V = 12$, $p = 1$	$V = 27$, $p = .56$	$V = 6$, $p = 1$
Peaches	$V = 19$, $p = 1$	$V = 21$, $p = .56$	$V = 7$, $p = 1$
Corn	$V = 3$, $p = 1$	$V = 6$, $p = 1$	$V = 0$, $p = 1$
Horseradish	$V = 1$, $p = 1$	$V = 6$, $p = 1$	$V = 2$, $p = 1$
Lobster	$V = 5$, $p = 1$	$V = 6$, $p = 1$	$V = 0$, $p = 1$
Blueberries	$V = 10$, $p = 1$	$V = 15$, $p = .64$	$V = 11$, $p = 1$
Cheese	$V = 0$, $p = .85$	$V = 21$, $p = .56$	$V = 7$, $p = 1$
Pork	$V = 9$, $p = 1$	$V = 4$, $p = 1$	$V = 9$, $p = 1$

Table 5.6: Differences between scores of the training and test phase, between allergic and fever, and the pre-scoring and test phase for the individual foods, Experiment 2. Red: used to be significant but no longer with the Holm adjustment, Blank: no significant difference in scores.

between an allergic reaction and a fever, thus resulting in participants scoring them similarly. It is also interesting to see that cue C did not score significantly different to fever and allergic, which is as we expect, as for cue C these two outcomes are both new outcomes.

Difference in score pre-test and test phase: As we argued in the introduction, the previously existing connections between food an an allergic reaction might influence the scoring during the experiment. We want to compare the pre-scores given in the training phase, to the scores given in the test phase. If there is no difference between the scores to the training and test phase, nor to the test phase and pre-scores, this could indicate that the previous knowledge might have influenced the results, as the scores seemed to have stayed on the same level as before the experiment.

In this section we will discuss the scores of the pre-scoring, training and test phase of the individual foods. In Appendix A we also compare the scores given to allergic and fever for these individual foods. Note that all the scores discussed here are to the individual foods instead of to groups of cues, as before.

First we looked at whether or not there was a difference between the scores to the individual foods in the original training phase (the last trial that cue was asked), and the test phase. In Table 5.6 we see that there was no difference for any of the foods (Training phase versus Test column). This also supports the notion that no learning took place during the test phase.

We investigated if the scores given in the test phase and in the pre-scoring differed. In the last column of Table 5.6 we can see that there are no significant differences between the scores. Thus the training and test phase were not scored differently, nor were the test and pre-scores. It could be that the pre-scoring might have had an influence in the whole experiment, which would explain the lack of a difference in both the test and training phase and the test and pre-scoring scores.

However it is important to note that not all participants saw each food, as the food was dependent on condition. This also meant that some food only had two data-points, so it is important to keep that in mind before drawing strong conclusions from these results.

5.2.3 Survey Results

The first question of the survey participants answered was "Did you find the instructions at the beginning of the experiment clear?". Participants could answer this open question with any amount of characters. The goal of this question was to see if we did indeed improve the instructions of the experiment compared to the previous experiment.

More than three quarters of the participants (78%) indicated that they understood the instructions, 11% was unsure if they understood them and 11% indicated that they did not understand at all. This seems to indicate that our instructions were indeed more clear as compared to our previous

experiment.

5.3 Discussion

This experiment aimed to perform an exact replication of the experiment Van Hamme and Wasserman describe in their 1994 paper. We wanted to replicate their findings, as well as disentangle which of the two model predictions would best describe the behaviour found in the test phase of our simulations.

If in the test phase the cues were scored equally as high as in their original block, this would be in line with the Rescorla-Wagner model. When the cues of the first block would be scored lower in the test phase as in the original blocks, and the cues of the third block scored equal, then this would be in line with the Van Hamme-Wasserman model. We also made predictions of the cues in relation to a newly introduced cue. The Rescorla-Wagner model predicted that the cues of the first and third block would be more predictive of an outcome than a new cue, while the Van Hamme-Wasserman model predicts that cues from the first block would be less predictive of an outcome than a new cue and the cues from the third block would be more predictive of an outcome.

We found support in the direction of the Rescorla-Wagner model, as the scores between the training and test phase did not differ, except for cues B and X of the first block on the overall scores. The scores to the new cue to allergic were lower for half of the cues in the overall scores, however when looking at the separate conditions there does not seem to be a clear pattern. The latter also supports the Rescorla-Wagner model, as for the Van Hamme-Wasserman model the new cue would have to be scored higher in the first block.

When comparing the results of our training phase to that of Van Hamme and Wasserman's results we find a similar kind of effect over condition, with the only difference being that their results are more differentiated. This can be found both in the average scoring for each condition, as for the scoring over trials. Participants seemed to understand how to score the trials better than the previous two experiments, as for almost all cues allergic trials were scored higher than not allergic trials and for those that were not scored differently a logical explanation was found. If we look at Figure 5.4 however, we can see that participants still score differently than expected, as in conditions 50-50 and 75-25, cues that we would expect to score higher, scored a 0 (definitely not leading to an allergic reaction) most often. This is slightly strange as one might expect that participants would want to score with caution and label something as "not safe to eat" if it could potentially cause an allergic reaction. What we saw here was the opposite, with participants answering that something is save to eat, even though it could cause an allergic reaction 50% of the time or even 75% of the time.

In this experiment, unlike Experiments 1 and 1B, we also asked participants to score the cues before they saw any of the trial's results. In this pre-score, we observed the following: there are preconceptions about which foods are more likely to cause an allergic reaction. While there were no cue *groups* (A, B or X) that were more likely to cause an allergic reaction according to the participants, there were certain individual foods such as peanuts, shrimp or walnuts. In our test phase we found that these scores had not really changed over the course of the experiment, which might indicate that the preconceived notions about the foods might influence the learning process. However it is important to note that there was little data for each individual food, so further research into this is still required.

One of the goals of this experiment was to increase the clarity of our results compared to the previous Experiment 1B. We found that the amount of participants understanding our instructions went from less than half to almost 80% in the current experiment. While too many factors were changed in this experiment to pin-point what exactly helped increase this clarity, we do think that clearly displaying the scale on screen increased clarity for rating quite a lot, as we now also saw more variability in scores in our frequency plots compared to the previous experiment.

As mentioned, we found that there was no difference between the scores of the cues in the pre-scoring and the scores of those same cues in the test phase. While it is important that a connection *can* be formed in the first place (see the Discussion of Experiment 1 (Garcia et al., 1968)), the fact that these connections have already been learned and updated over time, might have influenced the current learning process within the experiment. As discussed in the introduction, the food paradigm is a common method in EDL, however as M. Le Pelley et al. (2013) state in their paper, multiple researchers found that participants learn faster about cues they have seen predict something before, than cues that were not predictive, indicating that previous knowledge will indeed matter in the learning process.

It is also important to note that we did expect to find a connection formed between allergic and the cues, and thus for it to score higher than fever in most cases. This was not what we found. While in the previous two experiments we doubted if a connection was formed in the first place, we know that connections between the cues and outcomes existed in this experiment because of

the pre-scoring. This effect could be because fever and allergic might not have been distinct enough as two outcomes. Even though fever is not a symptom of an allergic reaction (symptoms are, amoungst others: swelling, difficulty breathing and rash: NHS, 2019b), participants might have believed it to be similar. In future experiments it might be useful to create an outcome that could be caused by foods, but that would be completely separate from an allergic reaction.

To conclude, we found evidence supporting the Rescorla-Wagner model and we replicated the findings of the experiment of Van Hamme and Wasserman, although we found slightly more around average scores than they did. We also found that we increased the clarity of our experiment compared to the last experiment.

6 General Discussion

The goal of this thesis was to see if people learn from absent cues. We looked at both Rescorla and Wagner's (1972) implementation of Error Driven Learning (EDL) and the implementation of Van Hamme and Wasserman (1994). The main difference between these two theories is that the first does not update predictions when cues are absent, while the latter does.

To answer the research question, we first modelled the experiment of Van Hamme and Wasserman (1994), in which they argued that their theory would be an improvement upon the theory of Rescorla and Wagner. The goal of creating a computational simulation was to find the different predictions that the two models would make. The results of the modelling showed that when we look at the relative weight (outcome - no outcome), both models predict approximately the same (albeit on a different scale). However if we look at the connections to *just* the outcome over the whole experiment, the differences between the two models become apparent. Cues that are no longer seen decrease rapidly in connection weight to both the outcome and the lack of an outcome in the Van Hamme-Wasserman model, while in the Rescorla-Wagner model the connection weights stay the same when the cues are no longer seen. However in their experiment, Van Hamme and Wasserman only asked about cues within their respective blocks, so they might not have found this distinction in their experiment setup. We also found that we could not tease apart the allergy and no allergy outcome. Therefore we needed new cues and outcomes to test the models.

To test the effect of decreasing weights over time, we added a simulated test phase to the modelling. In this test phase, cues from the first and last block were asked again, as well as a completely new cue. From this we predicted that in the Rescorla-Wagner model the cues' connection weights will be similar to the weights in their original block and that all cues would have a *higher* weight than a new cue only introduced in the test phase. For the Van Hamme-Wasserman model, our simulations showed that the connection weights have decreased in the test phase for the first block, but not for the third, compared to their original blocks. We also predicted that the cues from the first block would all have a *lower* weight than a new cue introduced in the test phase, while those of the third would have a higher weight than the new cue.

These models thus predict that, if the cues are asked again after their original training phase, the connection weights of the Rescorla-Wagner model are constant, while those of the Van Hamme-Wasserman model decrease in the first block. As we cannot see directly how strong the connections between the cues and outcomes are in participants, we expect the connection weights (or prediction strength) to be somewhat equal to the causal rating scores from the 1994 experiment. A higher connection weight would be equal to a higher causal rating score

To test our model predictions, we performed three experiments. In our first two experiments we replicated the experiment of Van Hamme and Wasserman (1994), but we adjusted the stimuli. We used treasure and objects instead of allergies and food. The training phase was the same as that of the 1994 experiment, but at the end of the experiments participants also saw a test phase where previous cues were asked again. In both experiments we replicated the results of Van Hamme and Wasserman, but the effects over conditions were much smaller. Participants often gave low causality ratings regardless of condition and they did not score outcome trials higher than no outcome trials for most cues.

In the first experiment we found support for the Rescorla-Wagner model when looking at the new cue and the difference between test and training phase, while in the second experiment we did find evidence for the Rescorla-Wagner model when looking at the differences between train and test, but not when looking at the new cue.

As both experiments did replicate the results of Van Hamme and Wasserman, but with a smaller effect of condition, together with the fact that the participants gave ratings that were not in line with our expectations, we did an exact replication of the 1994 experiment. We presented participants with food and asked them to rate the causality of the foods to an allergic reaction. While the training

phase was an exact replication, we kept the test phase in this experiment. In this experiment we replicated the results of Van Hamme and Wasserman, and we found evidence in the direction of the Rescorla-Wagner model in both the scores to the new cue and the scores to test and train.

Overall we found support for the predictions the Rescorla-Wagner model made in our computational modelling, therefore it seems that we do not learn in the absence of cues.

We also found that previously learned cues can influence the learning process, as we found no difference in the pre-scoring in our last experiment and the scores of the test phase, nor between the scores for a majority of the cues of the test and training phase. This is in line with findings from M. Le Pelley et al. (2013).

One other important practical implication that we found was that it is important that connections between the objects and outcomes can occur in the first place. As Garcia et al. (1968) showed as well, we found evidence that there might not have been a connection between the objects and the outcomes in the first two experiments, because participants did not believe a connection was possible in the first place. Having an outcome with a clear connection to the cues, such as food and allergies, does make sure that participants understand that a connection can be formed. This does lead to the problem that participants already have connections before the experiment and thus do not fully learn these connections in the experimental context.

It is important to bear in mind that the original experiment was done offline, while our current experiments were all done online. This was both due to lockdown regarding the Corona virus and because we were able to obtain a bigger sample size online than offline. However it is possible that this would cause a bias in the responses as it is more difficult to gauge if participants understood the experiment and were motivated to take part in it. M. E. Le Pelley et al. (2015) stated that participants will only engage in a controlled reasoning process, if they have the motivation and the opportunity to do so. As we could not control for the experimental environment we cannot be sure that these two factors were present.

One way to account for part of this would be to change the design of the experiment to involve implicit reasoning instead of explicit ratings, as this would shift the experiment away from a controlled reasoning process. As mentioned in the introduction, Reber (1989) and Ramscar, Dye, and Klein (2013) found that, explicit inference could hinder the process of implicit learning. Changing the experiment from an explicit reasoning task to an implicit one might help to engage participants more, but it could also reduce the amount of explicit reasoning that could interfere with the implicit process of EDL.

One more way in which an implicit task might change the performance is via the absent cue. In the current experiment setup, while what Van Hamme and Wasserman call the absent cue is not shown on screen, participants still have to score according to that cue. Thus one could argue that this cue is not truly absent, as a participant still sees the cue in some way. An experiment design that allows for implicit learning would allow us to create fully absent cues. The results of this design could then be compared to the one with explicit scoring, to see if the learning of these absent cues changes depending on the level off absentness.

As mentioned in the introduction, one of the original goals of this thesis was to create an experiment with this implicit learning design, yet due to time constraints and not fully replicating original results, we did not get to perform it. As argued above, we do however still believe that creating such an experiment would give more insight into learning and how we deal with absent cues.

7 Conclusion

This thesis set out to find if people learn from absent cues. Through several experiments we have found support for the Rescorla-Wagner predictions, which we obtained by modelling the experiment Van Hamme and Wasserman performed in 1994. This would thus indicate that people do not learn from absent cues. We also found that previously learned connections between cues and outcomes could influence the learning process within the experiment, but that it also important that a connection *is* possible in the first place.

The current research focused on explicitly given ratings for the causality of an outcome resulting from a cue, which might have influenced the (implicit) learning process of EDL. Future research could investigate this further by introducing a similar experiment but enforcing an implicit experimental design, thus creating an absent cue that is fully absent.

8 Acknowledgements

I would like to thank Dorothée Hoppe for helping me understand how to model Error Driven Learning in R and explaining the process of running experiments on Prolific. I also wish to show my

appreciation for my supervisors Jessie Nixon an Jacolien van Rij (and later Jelmer Borst) for giving me great feedback and guiding me throughout this project. I would also like to thank Peter for being the world's greatest rubber duck, and for helping me debug a lot of my code.

References

Arnold, D., Tomaschek, F., Sering, K., Lopez, F., & Baayen, R. H. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PloS one*, *12*(4).

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied linguistics*, *27*(1), 1–24.

Garcia, J., McGowan, B. K., Ervin, F. R., & Koelling, R. A. (1968). Cues: Their relative effectiveness as a function of the reinforcer. *Science*, *160*(3829), 794–795.

Gelb, M. J., & Buzan, T. (1996). *Lessons from the art of juggling: How to achieve your full potential in business, learning, and life*. Three Rivers Press.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70.

Hoppe, D. B., Hendriks, P., Ramscar, M., & van Rij, J. (2020). An exploration of error-driven learning in simple two-layer networks from a discriminative learning perspective.

Houwer, J. D., & Beckers, T. (2002). A review of recent developments in research and theories on human contingency learning. *The Quarterly Journal of Experimental Psychology: Section B*, *55*(4), 289–310.

Hsu, A. S., Chater, N., & Vitányi, P. M. (2011). The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis. *Cognition*, *120*(3), 380–390.

Kamin, L. J. (1967). Attention-like processes in classical conditioning.

Lassaline, M. E., Wisniewski, E. J., & Medin, D. L. (1992). 9 basic levels in artificial and natural categories: Are all basic levels created equal? In *Advances in psychology* (Vol. 93, pp. 327–378). Elsevier.

Le Pelley, M., Calvini, G., & Spears, R. (2013). Learned predictiveness influences automatic evaluations in human contingency learning. *Quarterly Journal of Experimental Psychology*, *66*(2), 217–228.

Le Pelley, M. E., Pearson, D., Griffiths, O., & Beesley, T. (2015). When goals conflict with values: counterproductive attentional and oculomotor capture by reward-related stimuli. *Journal of Experimental Psychology: General*, *144*(1), 158.

Markman, A. B. (1989). Lms rules and the inverse base-rate effect: Comment on gluck and bower (1988).

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). Opensesame: An open-source, graphical experiment builder for the social sciences. *Behavior research methods*, *44*(2), 314–324.

NHS. (2019a). Food allergy - causes - nhs. Retrieved April 2021, from `https://www.nhs.uk/conditions/food-allergy/causes/`

NHS. (2019b). Food allergy - symptoms - nhs. Retrieved April 2021, from `https://www.nhs.uk/conditions/food-allergy/symptoms/`

Nixon, J. S. (2020). Of mice and men: Speech sound acquisition as discriminative learning from prediction error, not just statistical tracking. *Cognition*, *197*, 104081.

Nixon, J. S., & Tomaschek, F. (2020). Learning from the acoustic signal: Error-driven learning of low-level acoustics discriminates vowel and consonant pairs. In *Proceedings of the 42nd annual meeting of the cognitive science society* (pp. 585–591).

Palan, S., & Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ramscar, M., Dye, M., Gustafson, J., & Klein, J. (2013). Dual routes to cognitive flexibility: Learning and response-conflict resolution in the dimensional change card sort task. *Child development*, *84*(4), 1308–1323.

Ramscar, M., Dye, M., & Klein, J. (2013). Children value informativity over logic in word learning. *Psychological science*, *24*(6), 1017–1023.

Ramscar, M., Sun, C. C., Hendrix, P., & Baayen, H. (2017). The mismeasurement of mind: Life-span changes in paired-associate-learning scores reflect the "cost" of learning, not cognitive decline. *Psychological science*, *28*(8), 1171–1179.

Ramscar, M., Thorpe, K., & Denny, K. (2007). Surprise in the learning of color words. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 29).

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of experimental psychology: General*, *118*(3), 219.

Rescorla, R. A. (1972). Informational variables in pavlovian conditioning. In *Psychology of learning and motivation* (Vol. 6, pp. 1–46). Elsevier.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American psychologist*, *43*(3), 151.

Rescorla, R. A., & Wagner, A. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning ii*, *64*, 99.

St. Clair, M. C., Monaghan, P., & Ramscar, M. (2009). Relationships between language structure and language learning: The suffixing preference and grammatical categorization. *Cognitive Science*, *33*(7), 1317–1329.

Tassoni, C. J. (1995). The least mean squares network with information coding: A model of cue learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 193.

van Rij, J. (2020). plotfunctions: Various functions to facilitate visualization of data and analysis [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=plotfunctions` (R package version 1.4)

van Rij, J., & Hoppe, D. (2020). edl: Toolbox for error-driven learning simulations with two-layer networks [Computer software manual]. (R package version 0.3)

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and motivation*, *25*(2), 127–151.

Widrow, B., & Hoff, M. E. (1960). *Adaptive switching circuits* (Tech. Rep.). Stanford Univ Ca Stanford Electronics Labs.

	Diamond	Fossil
A1	$t(58) = -0.61$, $p = 1$	$t(58) = -1.26$, $p = 1$
B1	$t(58) = 0.54$, $p = 1$	$t(58) = -2.42$, $p = .13$
X1	$t(59) = -1.08$, $p = 1$	$t(59) < 0.001$, $p = 1$
C	$t(58) = 0.58$, $p = 1$	$t(58) = -0.25$, $p = 1$
A3	$t(59) = -0.67$, $p = 1$	$t(59) = 2.04$, $p = .28$
B3	$t(59) = 0.11$, $p = 1$	$t(59) = -0.57$, $p = 1$
X3	$t(59) = -0.24$, $p = 1$	$t(59) = 1.21$, $p = 1$

Table A.1: Differences between part 1 and part 2 of the test phase for the outcomes diamond and fossil, Experiment 1. Red: used to be significant but no longer with the Holm adjustment, Blank: no significant difference in scores.

A Appendix A; Results supplements

In this appendix we will discuss the results that were found in our experiments, but that were not used to answer the main question of this thesis.

A.1 Experiment 1, results

We will first look at the differences between part 1 and 2 of the test phase, if there were any differences in scores when looking over conditions and lastly we will discuss the third part of the test phase.

A.1.1 Differences between part 1 and 2 of the test phase

As already mentioned in the results section, there were no significant differences between the scores given to part 1 and part 2 of the test phase. This was neither for the scores to diamond (the already learned outcome) and the new outcome fossil. The results of the paired t-tests with Bonferroni-Holm correction for the p-values can be seen in Table A.1.

A.1.2 Differences between cue scores between conditions

While we did look at the effect of the different conditions for the main two parts of our hypothesis for the test phase (training versus test scores and a new cue), we also investigated if there was a difference in score in the test phase of a cue to diamond for the different conditions. The results of this can be seen in Table A.2. There were no differences in scores between conditions for Experiment 1.

A.1.3 Test phase part 3, cues from the same category

In this last part of the test phase, participants saw cues that they had not seen before. These cues belonged to the same category, as the cues that they saw in the training phase. This part of the test phase was done mostly as a reference for future experiment setups. We assumed that participants would not only have created a connection between a cue and an outcome, but also between the features of that cue's category and the outcome. If this was indeed the case we expect to see no differences between the score given in this test phase and the score given to the original cue in the training phase.

Table A.3 contains the results of the Wilcoxon tests and t-tests done for this part of the test phase. In the first column are the results of the Wilcoxon test between the scores of the original cues and their alternative counterpart, the second column shows the results of the t-tests between the scores to diamond in the test phase and no diamond in the test phase.

Except for cue A from the first block, where the original cue scored higher than the alternative version in the test phase, there were no significant differences between scores. While it might seem that this means that participants recognised that these cues belonged to the same category, all the scores lie around 4 to 5, which might also indicate that people were not very sure how to score these new cues. This interpretation is further supported by the fact that there were no significant differences between the scores to the diamond outcome in the test phase and the scores to the no diamond outcome in the test phase, which *was* a difference that was present at least in cues A and X in the training phase.

	50-50 vs 75-25	75-25 vs 100-0	100-0 vs 50-50
A1	$V = 215$, $p = 1$	$V = 175$, $p = 1$	$V = 206$, $p = 1$
B1	$V = 222$, $p = 1$	$V = 159$, $p = 1$	$V = 228$, $p = 1$
X1	$V = 220$, $p = 1$	$V = 185$, $p = 1$	$V = 214$, $p = 1$
A3	$V = 174$, $p = 1$	$V = 280$, $p = 1$	$V = 146$, $p = 1$
B3	$V = 119$, $p = 1$	$V = 276$, $p = 1$	$V = 209$, $p = 1$
X3	$V = 134$, $p = 1$	$V = 274$, $p = 1$	$V = 200$, $p = 1$

Table A.2: Differences in scores to diamond between conditions for Experiment 1. Blank: no significant difference in scores.

	Original and alternative	**Diamond and No Diamond**
A1	$V = 416$, $p = .02$	$t(59) = -0.68$, $p = 1$
B1	$V = 729$, $p = .76$	$t(59) = -0.19$, $p = 1$
X1	$V = 874$, $p = .76$	$t(59) = 0.66$, $p = 1$
A3	$V = 988$, $p = .34$	$t(59) = 1.30$, $p = 1$
B3	$V = 850$, $p = .57$	$t(59) = -0.04$, $p = 1$
X3	$V = 849$, $p = .57$	$t(59) = -0.15$, $p = 1$

Table A.3: Differences between scores of the original and alternative cues, and the difference between the scores to diamond and no diamond of the alternative cues, Experiment 1. Green: Original higher than alternative, Blank: no significant difference in scores.

	Diamond	Fossil
A1	$t(9) = -1.67$, $p = .90$	$t(9) = -2.33$, $p = .27$
B1	$t(9) = -0.69$, $p = 1$	$t(9) = -1.31$, $p = .67$
X1	$t(9) < 0.001$, $p = 1$	$t(9) = -1.70$, $p = .61$
C	$t(9) = -1.63$, $p = .90$	$t(9) = 1.63$, $p = .61$
A3	$t(9) = 0.80$, $p = 1$	$t(9) = 0$, $p = 1$
B3	$t(9) = 0.45$, $p = .26$	$t(9) = -2.45$, $p = .26$
X3	$t(9) = 0.45$, $p = .67$	$t(9) = 1.11$, $p = .67$

Table A.4: Differences between part 1 and part 2 of the test phase for the outcomes diamond and fossil, Experiment 1B. Red: used to be significant but no longer with the Holm adjustment, Blank: no significant difference in scores.

A.2 Experiment 1B, results

We will first look at the differences between part 1 and 2 of the test phase, if there were any differences in scores when looking over conditions and lastly we will discuss the third part of the test phase.

A.2.1 Differences between part 1 and 2 of the test phase

As already mentioned in the results section, there were no significant differences between the scores given to part 1 and part 2 of the test phase. This was neither for the scores to diamond (the already learned outcome) and the new outcome fossil. The results of the paired t-tests with Bonferroni-Holm correction for the p-values can be seen in Table A.4.

A.2.2 Differences between cue scores between conditions

While we did look at the effect of the different conditions for the main two parts of our hypothesis for the test phase (training versus test scores and a new cue), we also investigated if there was a difference in score of a cue to diamond for the different conditions. The results of this can be seen in Table A.5. There were no differences in scores between conditions for Experiment 1B.

A.2.3 Test phase part 3, cues from the same category

In this last part of the test phase, participants saw cues that they had not seen before. These cues belonged in the same category however, as the cues that they saw in the training phase. This part of the test phase was done mostly as a reference for future experiment setups. We hoped that participants would not only have created a connection between a cue and an outcome, but also between the features of that cue's category and the outcome. If this was indeed the case we expect to see no differences between the score given in this test phase and the score given to the original cue in the training phase.

Table A.6 contains the results of the Wilcoxon tests done for this part of the test phase. In the first column are the results of the Wilcoxon test between the scores of the original cues and their alternative counterpart, the second column shows the results of the Wilcoxon tests between the scores to diamond in the test phase and no diamond in the test phase.

There were no significant differences between

	50-50 vs 75-25	75-25 vs 100-0	100-0 vs 50-50
A1	$V = 12$, $p = 1$	$V = 4$, $p = 1$	$V = 2$, $p = 1$
B1	$V = 8$, $p = 1$	$V = 3$, $p = 1$	$V = 4$, $p = 1$
X1	$V = 6$, $p = 1$	$V = 5$, $p = 1$	$V = 3$, $p = 1$
A3	$V = 5$, $p = 1$	$V = 5$, $p = 1$	$V = 5$, $p = 1$
B3	$V = 2$, $p = 1$	$V = 12$, $p = 1$	$V = 6$, $p = 1$
X3	$V = 7$, $p = 1$	$V = 5$, $p = 1$	V = 3, p = 1

Table A.5: Differences in scores to diamond between conditions for Experiment 1B. Blank: no significant difference in scores.

	Original and alternative	**Diamond and No Diamond**
A1	$V = 5$, $p = .15$	$V = 8$, $p = 1$
B1	$V = 22$, $p = 1$	$V = 11$, $p = 1$
X1	$V = 26$, $p = 1$	$V = 14$, $p = 1$
A3	$V = 2$, $p = .15$	$V = 28$, $p = 1$
B3	$V = 17$, $p = 1$	$V = 12$, $p = 1$
X3	$V = 11$, $p = .68$	$V = 9$, $p = 1$

Table A.6: Differences between scores of the original and alternative cues, and the difference between the scores to diamond and no diamond of the alternative cues, Experiment 1B. Red: used to be significant but no longer with the Holm adjustment, Blank: no significant difference in scores.

scores. While it might seem that this means that participants recognised that these cues belonged to the same category, all the scores lie around 4 to 5, which might also indicate that people were not very sure how to score these new cues. This interpretation is further supported by the fact that there were no significant differences between the scores to the diamond outcome in the test phase and the scores to the no diamond outcome in the test phase, which *was* a difference that was present at least in cues A and X in the training phase.

	Allergic	Fever
A1	$t(18) = -0.25$, $p = 1$	$t(18) = 0.43$, $p = 1$
B1	$t(18) = -0.29$, $p = 1$	$t(18) = -0.81$, $p = .13$
X1	$t(18) = -0.56$, $p = 1$	$t(18) = -0.34$, $p = 1$
C	$t(18) = -0.27$, $p = 1$	$t(18) = 0.35$, $p = 1$
A3	$t(18) = 2.45$, $p = .17$	$t(18) = -1.33$, $p = .28$
B3	$t(18) = 0.63$, $p = 1$	$t(18) = -0.53$, $p = 1$
X3	$t(18) = 1.66$, $p = .37$	$t(18) = -0.93$, $p = 1$

Table A.7: Differences between part 1 and part 2 of the test phase for the outcomes allergic and fever, Experiment 2. Red: used to be significant but no longer with the Holm adjustment, Blank: no significant difference in scores.

A.3 Experiment 2, results

We will first look at the differences between part 1 and 2 of the test phase, if there were any differences in scores when looking over conditions, and lastly we will discuss the third and fourth part of the test phase.

A.3.1 Differences between part 1 and 2 of the test phase

As already mentioned in the results section, there were no significant differences between the scores given to part 1 and part 2 of the test phase. This was neither for the scores to allergic reaction (the already learned outcome) and the new outcome fever. The results of the paired t-tests with Bonferroni-Holm correction for the p-values can be seen in Table A.7.

	50-50 vs 75-25	75-25 vs 100-0	100-0 vs 50-50
A1	$V = 11$, $p = 1$	$V = 2$, $p = .29$	$V = 48$, $p = .04$
B1	$V = 13$, $p = 1$	$V = 15$, $p = 1$	$V = 35$, $p = 1$
X1	$V = 20$, $p = 1$	$V = 10$, $p = 1$	$V = 32$, $p = 1$
A3	$V = 40$, $p = .11$	$V = 10$, $p = 1$	V = 12, $p = 1$
B3	$V = 22$, $p = 1$	$V = 12$, $p = 1$	$V = 28$, $p = 1$
X3	$V = 14$, $p = 1$	$V = 31$, $p = .65$	$V = 16$, $p = 1$

Table A.8: Differences in scores to allergic between conditions for Experiment 2. Red: used to be significant but no longer with the Holm adjustment, Green: former condition scored higher than latter, Blank: no significant difference.

	Original and alternative	**Allergic and not allergic**
A1	$V = 52$, $p = .25$	$V = 93$, $p = .41$
B1	$V = 119$, $p = .04$	$V = 34$, $p = 1$
X1	$V = 146$, $p = .006$	$V = 82$, $p = 1$
A3	$V = 33$, $p = .07$	$V = 32$, $p = 1$
B3	$V = 110$, $p = .07$	$V = 25$, $p = 1$
X3	$V = 157$, $p = .06$	$V = 70$, $p = 1$

Table A.9: Differences between scores of the original and alternative cues, and the difference between the scores to allergic and not allergic of the alternative cues, Experiment 2. Red: used to be significant but no longer with the Holm adjustment, Green: Original higher than alternative, Blank: no significant difference in scores.

	Training phase versus Test	Allergy versus Fever	Pre-scoring versus Test
Shrimp	$V = 28$, $p = 1$	$V = 19$, $p = 1$	$V = 20$, $p = 1$
Strawberries	$V = 11$, $p = 1$	$V = 21$, $p = .56$	$V = 3$, $p = 1$
Peanuts	$V = 34$, $p = .60$	$V = 26$, $p = .61$	$V = 13$, $p = 1$
Yogurt	$V = 21$, $p = .60$	$V = 21$, $p = .56$	$V = 15$, $p = 1$
Bran	$V = 9$, $p = 1$	$V = 13$, $p = 1$	$V = 14$, $p = 1$
Cabbage	$V = 11$, $p = 1$	$V = 8$, $p = 1$	$V = 4$, $p = 1$
Bananas	$V = 13$, $p = 1$	$V = 6$, $p = 1$	$V = 0$, $p = 1$
Chicken	$V = 2$, $p = 1$	$V = 3$, $p = 1$	$V = 10$, $p = 1$
Mustard	$V = 7$, $p = 1$	$V = 3$, $p = 1$	$V = 4$, $p = 1$
Wheat	$V = 28$, $p = .28$	$V = 28$, $p = .40$	$V = 13$, $p = 1$
Walnuts	$V = 12$, $p = 1$	$V = 27$, $p = .56$	$V = 6$, $p = 1$
Peaches	$V = 19$, $p = 1$	$V = 21$, $p = .56$	$V = 7$, $p = 1$
Corn	$V = 3$, $p = 1$	$V = 6$, $p = 1$	$V = 0$, $p = 1$
Horseradish	$V = 1$, $p = 1$	$V = 6$, $p = 1$	$V = 2$, $p = 1$
Lobster	$V = 5$, $p = 1$	$V = 6$, $p = 1$	$V = 0$, $p = 1$
Blueberries	$V = 10$, $p = 1$	$V = 15$, $p = .64$	$V = 11$, $p = 1$
Cheese	$V = 0$, $p = .85$	$V = 21$, $p = .56$	$V = 7$, $p = 1$
Pork	$V = 9$, $p = 1$	$V = 4$, $p = 1$	$V = 9$, $p = 1$

Table A.10: Differences between scores of the training and test phase, between allergic and fever, and the pre-scoring and test phase for the individual foods, Experiment 2. Red: used to be significant but no longer with the Holm adjustment, Blank: no significant difference in scores.

A.3.2 Differences between cue scores between conditions

While we did look at the effect of the different conditions for the main two parts of our hypothesis for the test phase (training versus test scores and a new cue), we also investigated if there was a difference in score of a cue to allergic for the different conditions. The results of this can be seen in Table A.8. There were no significant differences in scores between conditions, except for cue A from the first block when comparing condition 100-0 and 50-50. Condition 100-0 was scored higher than 50-50, which is in line with expectations, as in condition 100-0 AX predicts an allergic reaction much more often as in condition 50-50. However this would not explain why only cue A from the first block is different, and not cue A from the third block.

A.3.3 Test phase part 3, cues from the same category

In part three of the test phase, participants saw cues that they had not seen before. These cues belonged in the same category however, as the cues that they saw in the training phase. This part of the test phase was done mostly as a reference for future experiment setups. We hoped that participants would not only have created a connection between a cue and an outcome, but also between the features of that cue's category and the outcome. If this was indeed the case, we expect to see no differences between the score given in this test phase and the score given to the original cue in the training phase.

Table A.9 contains the results of the Wilcoxon tests. In the first column are the results of the Wilcoxon test between the scores of the original cues and their alternative counterpart, the second column shows the results of the tests between the scores to allergic in the test phase and not allergic in the test phase.

For cues B and X from the first block, the original cue scored higher than the alternative version in the test phase. However for the majority of the cues there were no significant differences between the scores. While it might seem that this means that participants recognised that these cues belonged to the same category, if we look at the allergic and not allergic column we can see that there was no significant difference between these two types of trials. This *was* different for most cues in the training phase, thus indicating that participants might have just not been sure on how to score these new cues, instead of recognising that they belonged to the same category.

A.3.4 Test phase part 4, pre-scores and individual food items

This last part of the test phase was unique to this experiment. Participants were shown one food item and then had to indicate the causality score for both fever and allergic. In this manner we could compare the pre-scores given in the training phase, but also look at the connections of single cues instead of compound cues.

We already looked at the differences between train and test phase and pre-scoring and test phase in the main results section, here we will look at the differences between the individual food scores to fever and allergic

We investigated if the score given to allergy for a food was higher than that of fever. If these are not scored differently, there is either not a strong enough relationship created between allergy and the cue in the training phase, or fever and allergy are too similar as outcomes and are therefore seen as overlapping. In column Allergy versus Fever of Table A.10 we see that none of the foods differed in their scoring to allergy and fever. This is in line with what we found in the main results section, where we also did not find a difference in scores for allergy and fever.

As already mentioned in the main results section however, it is important to note that not all participants saw each food, as the food was dependent on condition. This also meant that some food only had two data-points, so it is important to keep that in mind when looking at these results and drawing strong conclusions from them.

B Appendix B; Figures and Tables

B.1 Computational Simulations

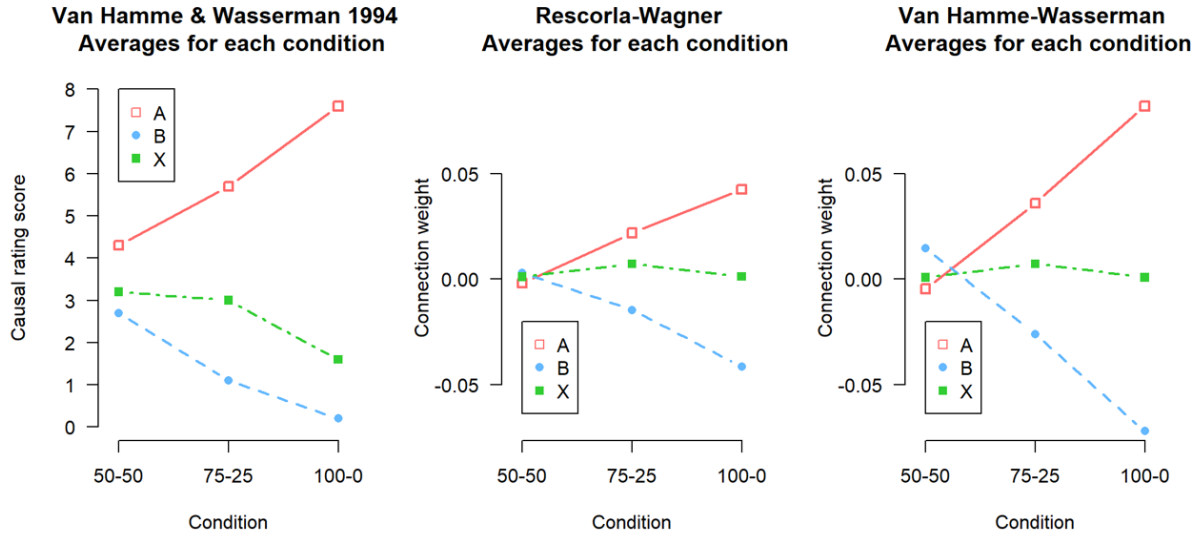

Figure B.1: Weights to allergic minus not allergic. Middle: Rescorla-Wagner model. Right: Van Hamme-Wasserman model. Left: Results of the original paper. Average for each condition.

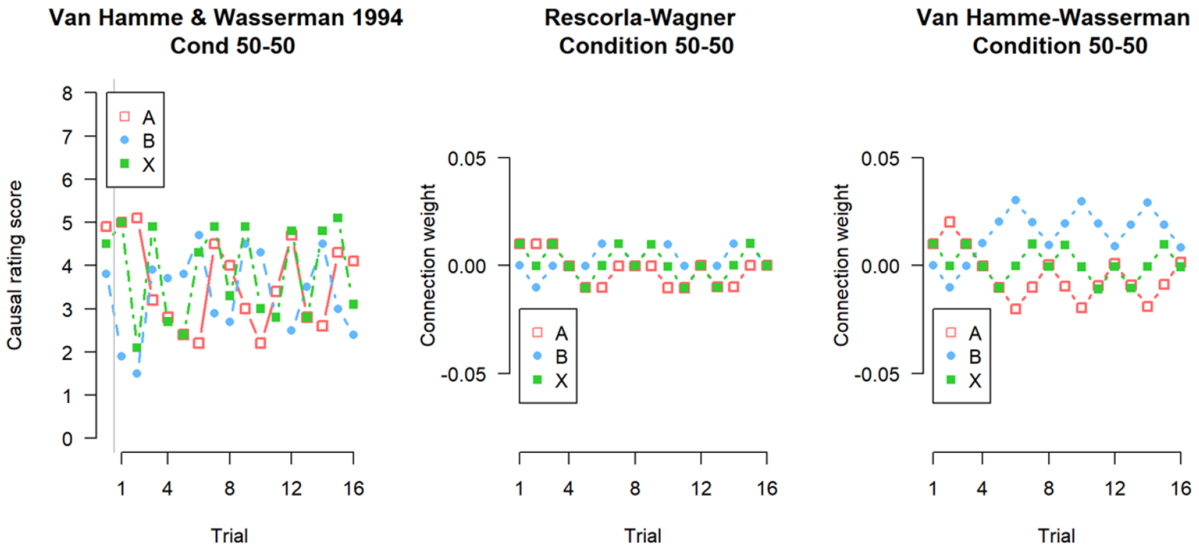

Figure B.2: Weights to allergic minus not allergic. Middle: Rescorla-Wagner model. Right: Van Hamme-Wasserman model. Left: Results of the original paper. Over trial for the 50-50 condition.

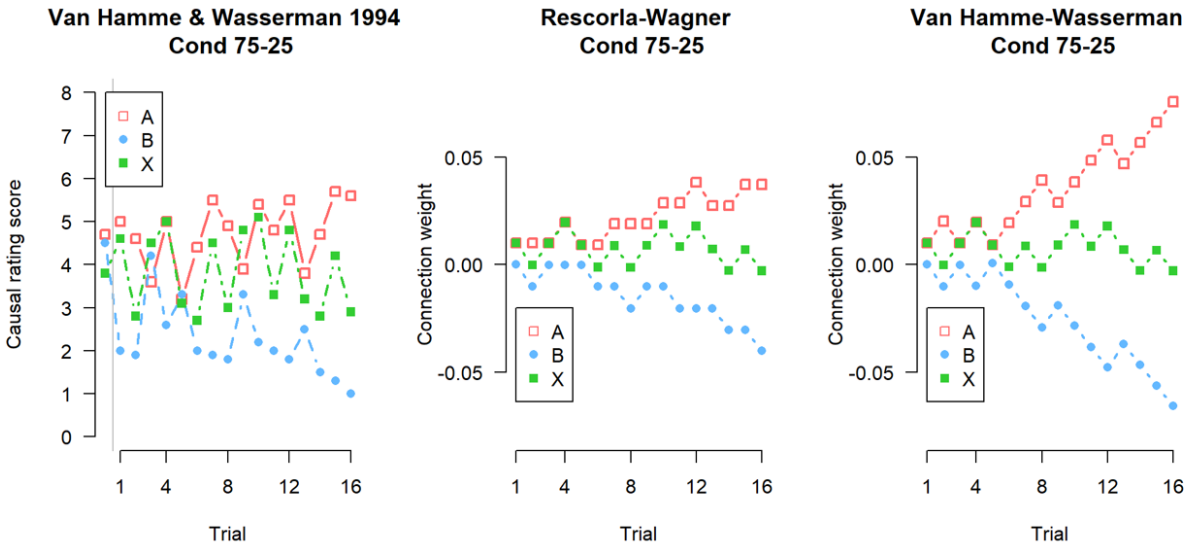

Figure B.3: Weights to allergic minus not allergic. Middle: Rescorla-Wagner model. Right: Van Hamme-Wasserman model. Left: Results of the original paper. Over trial for the 75-25 condition.

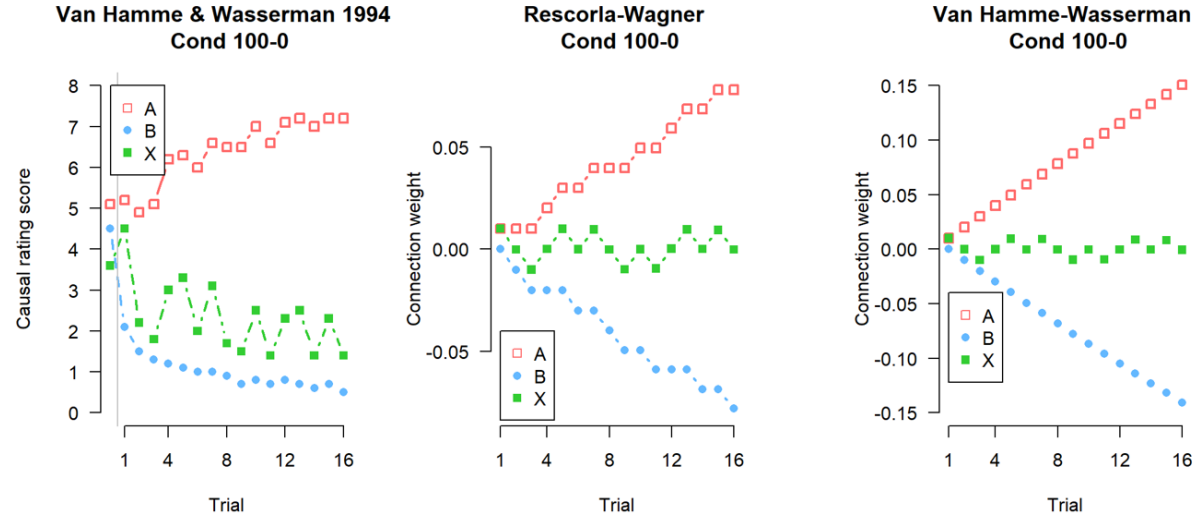

Figure B.4: Weights to allergic minus not allergic. Middle: Rescorla-Wagner model. Right: Van Hamme-Wasserman model. Left: Results of the original paper. Over trial for the 100-0 condition.

B.2 Experiment 1B

B.3 Questions in the Survey

- Q0: Please fill in your ProlificID, such that we can compare the survey to the test results.
- Q1: Did you find the instructions at the beginning of the experiment clear? If not, what was unclear?
- Q2: What do you think the experiment was about?
- Q3: What did you rate according to? What things did you take into account in your decision?
- Q4: How did you decide on your rating scores? How did you select the numbers?
- Q5: Did your rating strategy change throughout the experiment? If so in what way?
- Q6: Any other comments?

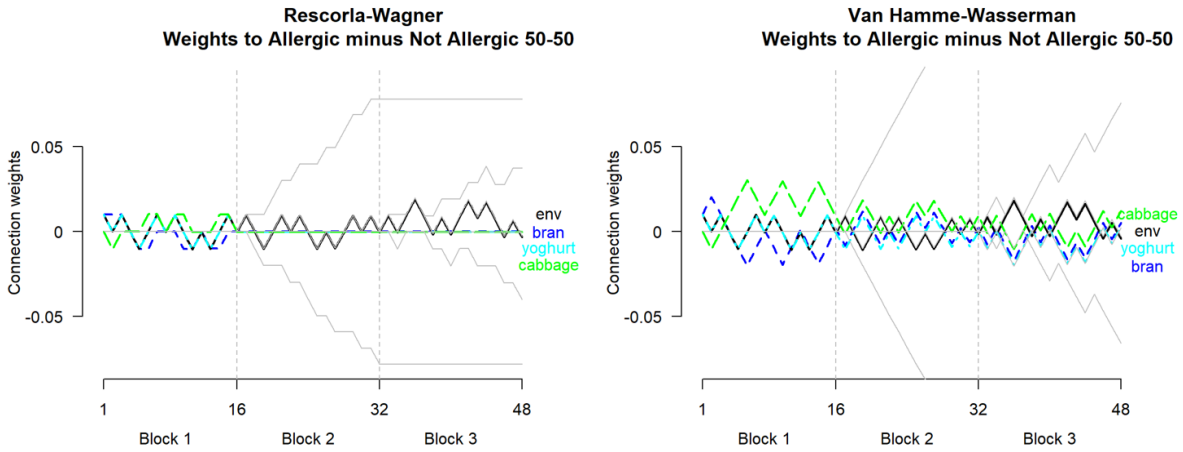

Figure B.5: Weights to allergic minus not allergic for each of the foods asked condition 50-50. Left: Rescorla-Wagner model. Right: Van Hamme-Wasserman model.

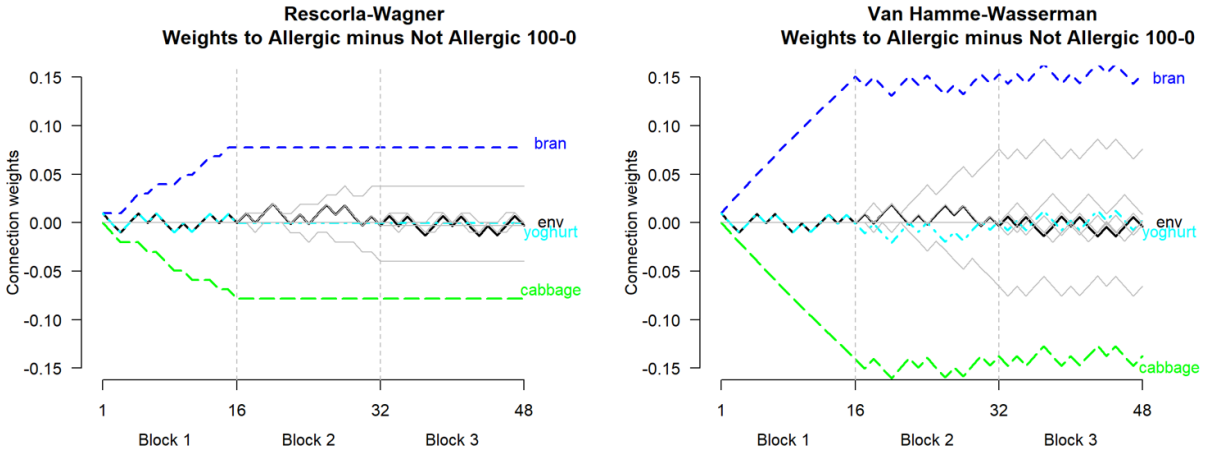

Figure B.6: Weights to allergic minus not allergic for each of the foods asked condition 100-0. Left: Rescorla-Wagner model. Right: Van Hamme-Wasserman model.

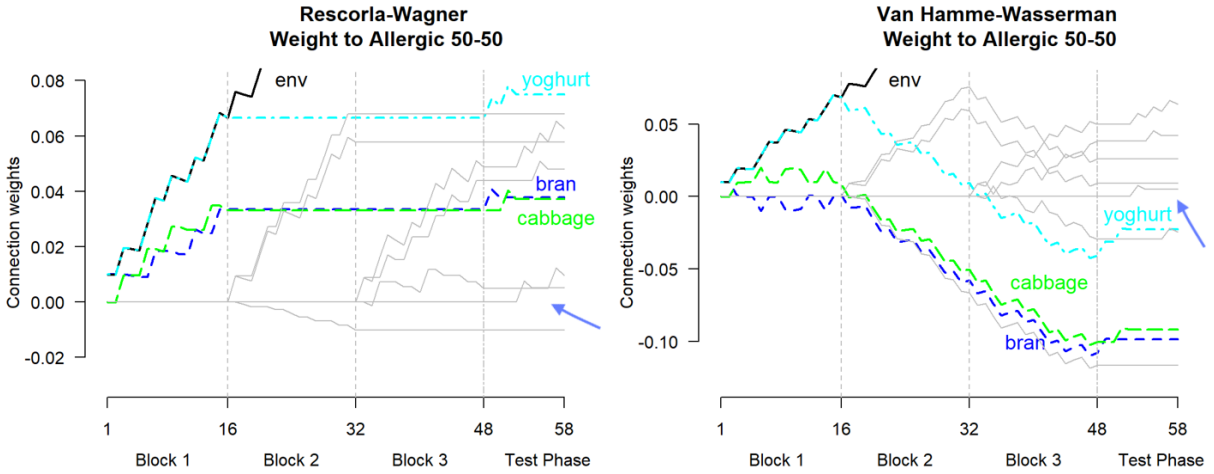

Figure B.7: Weights to allergic, with test phase, block 1 is condition 50-50. Arrows indicate the novel cue that is only introduced in the test phase. Left: Rescorla-Wagner model. Right: Van Hamme-Wasserman model.

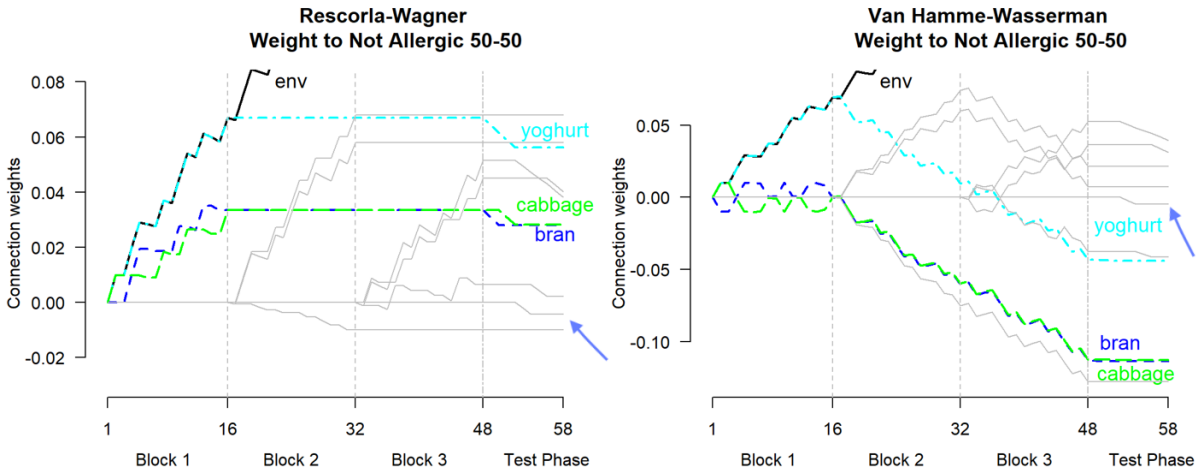

Figure B.8: Weights to not allergic, with test phase, block 1 is condition 50-50. Arrows indicate the novel cue that is only introduced in the test phase. Left: Rescorla-Wagner model. Right: Van Hamme-Wasserman model.

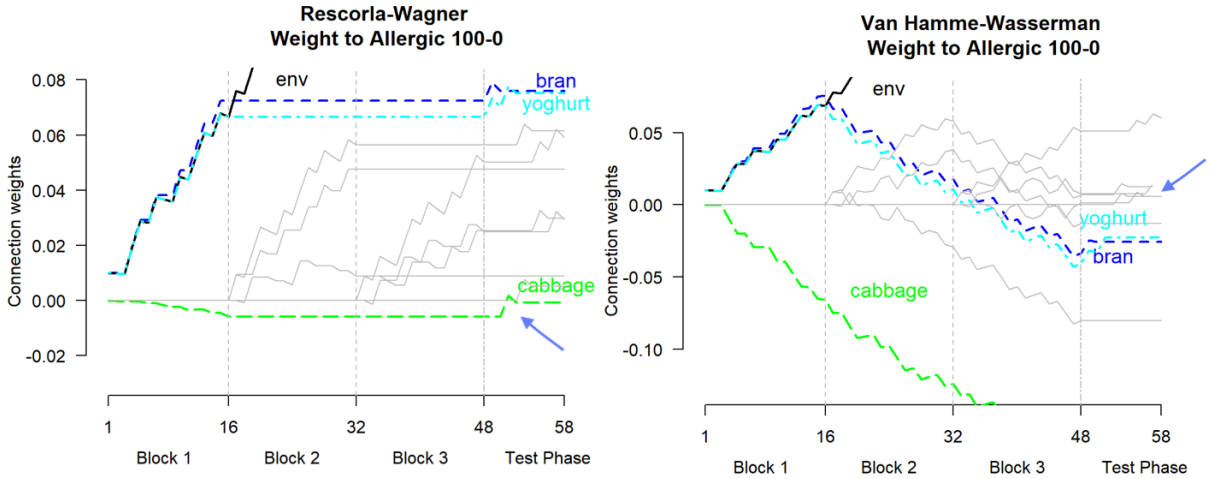

Figure B.9: Weights to allergic, with test phase, block 1 is condition 100-0. Arrows indicate the novel cue that is only introduced in the test phase. Left: Rescorla-Wagner model. Right: Van Hamme-Wasserman model.

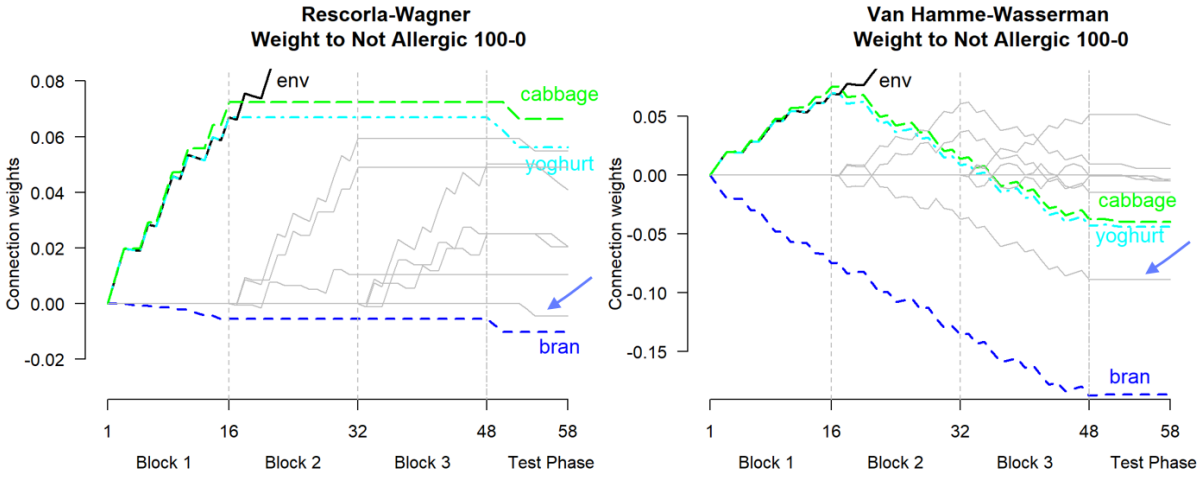

Figure B.10: Weights to not allergic, with test phase, block 1 is condition 100-0. Arrows indicate the novel cue that is only introduced in the test phase. Left: Rescorla-Wagner model. Right: Van Hamme-Wasserman model.

Group	Nr of Subjects	Condition Order	Food Group Order
1	7	75-25; 50-50; 100-0	2, 1, 3
		75-25; 100-0; 50-50	5, 6, 4
2	8	50-50; 100-0; 75-25	1, 3, 2
		50-50; 75-25; 100-0	4, 5, 6
3	6	100-0; 75-25; 50-50	3, 2, 1
		100-0; 50-50; 75-25	6, 4, 5
4	8	75-25; 100-0; 50-50	5, 6, 4
		75-25; 50-50; 100-0	2, 1, 3
5	10	50-50; 75-25; 100-0	4, 5, 6
		50-50;100-0; 75-25	1, 3, 2
6	9	100-0; 50-50; 75-25	6, 4, 5
		100-0; 75-25;50-50	3, 2, 1

Table B.1: Table 2 from the Van Hamme and Wasserman paper (1994), adjusted to our naming of the conditions. Note that we only used the first line of each group in our simulations and Experiment 2.

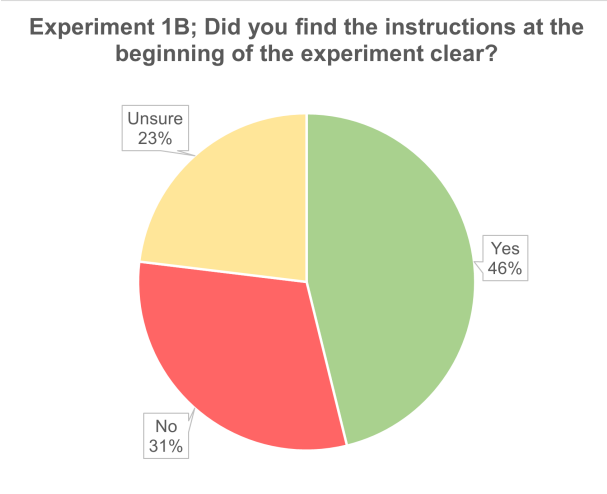

Figure B.11: Results of the survey question "Did you find the instructions at the beginning of the experiment clear?" from Experiment 1B.

B.4 Experiment 2

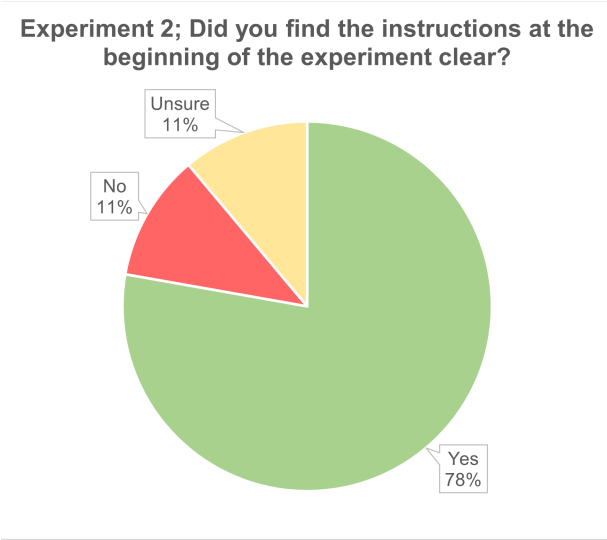

Figure B.12: Results of the survey question "Did you find the instructions at the beginning of the experiment clear?" in Experiment 2.