# UNDERSTANDING COMPLEX CONCEPTS

## INVESTIGATION OF THE OVEREXTENSION EFFECT USING THE TRUTH-VALUE JUDGMENT TASK

**MA THESIS**

**AUGUST 26TH, 2016**

**UTRECHT UNIVERSITY, FACULTY OF HUMANITIES**

**GENERAL LINGUISTICS: THE STUDY OF THE LANGUAGE FACULTY**

**SUPERVISOR PROF. DR. YOAD WINTER**

**2ND READER DR. RICK NOUWEN**

**ELETTA DAEMEN**

**STNR. 3389790**

**ABSTRACT**

This thesis is part of a larger scale project investigating how we understand complex concepts that are a conjunction of two simple categories (e.g. *food that is also a plant*). When we give our membership judgments in such complex categories, we often tend to overextend our linguistic categories: we agree to membership in the conjunction while disagreeing with membership in one of the constituent categories. This behavior is in contradiction with classical logic, that states that we can only agree with membership in the conjunction iff we agree with membership in both constituents. The effect that is described is widely known as the *overextension effect*.

This thesis offers an insight into overextension data for Dutch, using novel items and categories. The subjects of this study (n=63) were asked to give their membership judgments of an item (e.g. *bread*) in two simple categories (e.g. *food* and *plant*) and in their conjunction (*food that is also a plant*). Two different analyses of the data show that the presence of the overextension effect in our data is greatly dependent on the assumptions that underlie our analysis: responses inconsistent with classical logic are the effect of people changing their minds about their judgment in a constituent category (M1) OR inconsistent responses are due to noise (M2). Based on the outcomes of the current experiment, we are not able to choose between these models. Further investigation of the origin of inconsistent responses is needed so that we may take another look at how we understand complex concepts and what is the role of overextension in this process.

## CONTENTS

## 1. INTRODUCTION

At first sight, object classification seems to be an easy task. The meaning (and boundaries) of the categories we use in daily life seem to be clear: we know what *furniture*, *plants* and *food* are. Also, we have basic knowledge on clear members and non-members of these categories: *bread* is *food* but it is certainly not a piece of *furniture*, a chair is a piece of *furniture* but it is definitely not a *plant*, etc. Following classical sentential logic, we might assume that such membership judgments for simple concepts reflect sharp categorization using two possible truth values: true [+] or false [-].

This thesis will take a look at conjunctive complex concepts in particular. These complex concepts are a conjunction of two simple concepts, and can take multiple syntactic forms; relative clause (TOOLS THAT ARE ALSO WEAPONS), Adj-Noun (RED APPLE), Noun-Noun (BUS STOP). Several explanations have been put forward to explain how we understand complex categories. According to classic Boolean logic we understand such complex concepts as a combination of two simple concepts that are conjoined using the logical conjunction operator AND. Accordingly, our judgment about the conjunction is a logical combination of our judgments of the two simple concepts (see Figure 1).

| p | q | p ∧ q |
|---|---|---|
| T | T | T |
| T | F | F |
| F | T | F |
| F | F | F |

Figure 1: Truth table for conjunctions, classical logic. T=true=[+], F=false=[-].

For example, when we need to decide whether a DRILL is a TOOL THAT IS ALSO A WEAPON, we make our decisions on both categories separately, and apply the truth-table for the conjunctive operator to these two results. In this case, first we would give membership judgments to DRILL in the category TOOL [+] and to DRILL in the category WEAPON [-]. Following the truth table for conjunctions, membership to the conjunction has to receive a membership value of [-] since it prescribes we may only say yes to the conjunction iff we say yes to both constituents (see Figure 1). Conclusively, our judgment on the conjunctive category TOOLS THAT ARE ALSO WEAPONS should be [-] according to Boolean logic. These three responses in a triple of [tool,

weapon, and conjunction] can be visualized as [+--]. Responses that follow classical logic are [+++], [---], [+--] and [-+-].

Hampton (1988) investigated to what extent people satisfy Boolean logic when judging membership in complex concepts. In his experiment, he asked his participants to judge item membership in two simple categories: GAMES and SPORTS. Later, participants were asked to judge item membership in the complex category GAMES THAT ARE ALSO SPORTS. Interestingly, Hampton found that participants did not always follow Boolean logic. Even when they decided that an item was not a member of one of the simple categories, they would later still agree with its membership in the conjunctive complex category. For example, participants would judge that the item CHESS was a member of the category GAME, but not a member of the category SPORTS. Still, when judging item membership in the conjunction, participants would agree with membership in the complex category GAMES THAT ARE SPORTS [+]. These kind of responses do not follow logic and are what Hampton calls *overextensions*: [-++], [+-+] and [--+]. An alternative non-logic response is *underextension*. This response includes accepting membership in both constituent categories while rejecting membership in the conjunction: [++-]. Hampton showed that overextension appears significantly more than underextension, thus showing it is an effect.

This thesis aims at replicating Hampton's overextension results for Dutch. It is part of a larger-scale project that investigates the nature of the overextension effect (that is, what causes us to overextend our conjunctive complex categories?). However, to further investigate *the nature of the effect*, we first need to show that the overextension effect appears in Dutch. Hampton's (1988) experiment is replicated using newly constructed items and categories. Participants are asked to give their judgments on membership statements in two simple categories and in the conjunctive complex category. The items and categories used in this experiment can be adopted in a follow-up study, provided that they generate the overextension effect that we expect based on Hampton's findings.

In the following section we will discuss theories that have been proposed to explain the overextension phenomenon. We will first take a look at typicality and membership, two central notions in research regarding complex concepts. We will see that there are two different views on the overextension effect: the processing-compensation thesis (Chater, Lyon, & Myers, 1990) and the composite prototype hypothesis (Hampton, 1988). In the third section the experimental method of the current experiment is detailed, after which the results

are presented in the fourth section. The fifth section provides further analytical background, followed by a more in-depth analysis of the results. The sixth section of the thesis will interpret these results in light of theories of categorization and complex concepts. The concluding paragraph discusses what next steps should be undertaken in order to find an answer to the broader question of what causes inconsistent responses to appear for complex concepts.

## 2. THEORETICAL BACKGROUND

In the introduction we have seen that we can use membership judgments to investigate our understanding of simple and complex concepts. This section will provide a theoretical background for our experiment, discussing vagueness, fuzzy set theory and different accounts of the overextension effect.

### 2.1 MEMBERSHIP, FUZZINESS AND THE MIN RULE

Object categorization appears to be a simple, automatic process: we can effortlessly say that an apple is food and that a bird is not a weapon. Now, how do we make such categorization decisions? If we want to decide if the item TABLE falls in the category FURNITURE, we will need to decide on its *membership* in that category. How do we make such membership decisions? According to the Threshold Model (Hampton, 1995; Hampton, 2007), membership can be determined based on typicality ratings of an item in a certain category. In this sense, typicality is a measure of how representative a certain item is for a specific category. For example, ROBIN should get a high typicality rating in the category of BIRDS, while PENGUIN should get a low typicality rating for the category BIRDS. To know if an item is a member of a category, we must check if it is typical enough: does it reach the membership threshold for that category? Thus, if we want to know if a TABLE falls in the category FURNITURE, we check if its typicality rating reaches a certain threshold, so that it may be called a member of that category: (1) yes or (0) no.

However, if we look closer, category membership cannot always be decided on so easily; category boundaries can be rather unclear. For example, is a raisin a plant? In the early 1920's, Łukasiewicz started investigating such unclear cases, introducing a third option of undetermined truth value. This three-valued logic was later developed into "fuzzy" logic by Zadeh (1965), in which category boundaries are not fixed and may vary depending on various factors, including context. The category's exact meaning remains vague and may shift according to the object that is being classified. As a solution to the membership problem in such vague categories, Zadeh stated category membership as being graded. According to his fuzzy logic an object is placed on the interval [0,1] for any given simple category.

For the conjunction of concepts (*Is X an A that is also a B?*), Zadeh formulated the *min* rule: membership to any conjunction can be maximally as high as the minimum of its two constituent categories: I(a,b) = min{a,b}. For example, if the membership of BREAD in the category FOOD is .98, while BREAD in the category PLANT is .09, membership to the conjunction of these two categories may be maximally .09.

### 2.2 'THE MIN RULE IS WRONG'

Even though fuzzy sets are still quite acceptable in category membership research, the min rule was challenged by several effects in the past. In 1984, Smith and Osherson stated that 'the min rule is wrong' (1984: 340) since it cannot account for *the conjunction effect*, otherwise known as the *guppy effect*. This effect emerges when an item is more similar to a conjunction of concepts than to its constituents. It thanks its name to the classic example: a *guppy* is a better example of the conjunction *pet fish* than of the categories *pet* and *fish* separately. The effect was found for several items and categories for English (Smith & Osherson, 1984) and Dutch (Storms, De Boeck, Van Mechelen, & Ruts, 1998). Notably, the effect was never actually found for *guppies* and *pet fish*.

To test the conjunction effect, Smith and Osherson (1984) built a taxonomy of conjunctive concepts. This taxonomy is based on attribute and value information for each prototype of a given concept. For example, the prototype of the concept *apple* has the attribute-value pairs color-red, shape-round and texture-smooth (Smith, Osherson, Rips, & Keane, 1988). Each adjective-noun pair has its own diagnosticity value, indicating the probability of the noun being a good example of the concept if the adjective is true. For example, *red apple* is a positively diagnostic pair since the redness gives a positive indication of an item being an apple. Likewise, *brown apple* is a negatively diagnostic pair since the brownness gives negative indication of the item being an apple. *Unsliced apple* is an example of a nondiagnostic pair since slicedness does not give any indication of the item being an apple.

Smith and Osherson's tested the taxonomy in a typicality rating experiment in which participants were asked to rate pictures in adjective, noun and conjunction categories. The items that were pictured were either a *good match* or a *poor match* of the categories, indicating how well the pictured item was described by the adjective in the conjunction: showing an unsliced apple to be judged in the conjunction *unsliced apple* would be a good

match item, a sliced apple would be a poor match item for the same concept. The authors hypothesized that for the good match items, the conjunction effect should appear (thus: giving higher typicality rating to the conjunction than to both constituent categories), whereas it should not be present for poor match items. Two experiments showed that this was indeed the case, showing higher typicality ratings to the conjunction than to its constituents. The authors thus proved the conjunction effect for negatively diagnostic pairs (e.g. *brown apple*).

According to Smith and Osherson, these results show that Zadeh's min rule is wrong for typicality ratings. As an alternative, the authors present an explanation of the results in terms of prototype representations: for each judgment, we need to look at the similarity between the object and the representation of the concept. This representation includes various pieces of information of a concept, thus making it possible for us to form complex concepts. For example, for the concept *apple*, the sample representation includes (a) a set of relevant attributes (e.g. *shape*, *color*), (b) a set of possible values for each attribute (e.g. *red*, *brown* for the color attribute), (c) the most likely feature for each attribute (e.g. *red* for color, *round* for shape) and (d) the diagnosticity for each attribute (shape being more diagnostic than color in this case). The similarity between the object and the concept is calculated so that we can give the object a typicality rating for that concept.

Smith and Osherson have proposed a sophisticated way of calculating typicality ratings for complex concepts. According to the authors, such complex structures are not necessary for membership ratings, for these ratings do not go against fuzzy set theory. Membership judgments follow logical rules so that a creature is only a pet fish if it is both a pet and a fish. This lead the authors to conclude that typicality and membership are essentially different processes: typicality is based on *characteristic* features (and is graded) while membership is based on *defining* features (and is not graded).

### 2.3 THE OVEREXTENSION EFFECT

According to Hampton (1988), it is this last claim that causes the Osherson and Smith account to fail in fully explaining how we understand complex concepts. He believes that typicality and membership are above all based on the same underlying process: typicality is needed in order to determine membership. Thus, membership cannot be derived compositionally if typicality cannot be calculated compositionally. By conducting several experiments, Hampton

showed that classical logic does not only fail in predicting typicality values, but is also unsuccessful in predicting conjunction membership.

Hampton's research included an experiment that investigated how people judge the typicality and membership of conjunctive categories. A single pair of categories (*sports* and *games*) was tested against three category membership ratings: category A, category B and their conjunction AB. The experiment consisted of two phases, which had four weeks in between. In stage one, participants were asked to make membership judgments (yes or no) and rate all items accordingly on gameness and sportness on a -3 to +3 scale. In stage two the original subjects received another task: rating the same 55 items on 'sports which are games' or 'games which are sports'.

The experiment showed that participants were prone to allow category membership in the conjunction while rejecting category membership in one constituent category. This effect was present for all items and classes, but was most easily noticeable when items were good member of one class, but marginal to the other. These results show that membership judgments do not follow classical logic, contra Smith and Osherson (1984). This effect was later replicated by Chater et al. (1990) for items that are good member of one category and marginal to the other. A replication experiment was conducted, including other pairs of overlapping categories (e.g. *furniture-household appliances*). In this experiment, the two stages were separated by two weeks. Again, there was high amount of overextension: significantly more overextensions than underextensions, proving the overextension effect.

The overextension effect in Hampton's experiment is explained by his *composite prototype theory*. This is a more advanced version of prototype theory that accounts for composite concepts. When two categories are combined, some of their generic intensions are aggregated into a composite prototype; attributes of both constituent categories are merged. Importantly, the attributes that conflict are not transferred. Judging an item in the concept conjunction is then based upon similarity to the new prototype. Overextension is explained by this theory by higher similarity of the item to the composite prototype than to the constituent category prototypes.

## 2.5 THE COMPENSATION HYPOTHESIS

An alternative to Hampton's theory has been proposed by Chater et al (1990): the *compensation hypothesis*. The hypothesis states that membership judgment in conjunctive categories is the result of a best fit strategy: people are more lenient in their membership judgments when more factors are involved. When we are looking for a best fit for a list of attributes, we would usually also include an item that fails (slightly) on just one constituent. For example, when looking for a new apartment we may have a list of criteria it has to fulfill: it has to be affordable, in a central location, having a large surface area, etc. However, because there is such scarce of apartments that meet all of these criteria at once, we are likely to compromise on one of the criteria. Such compromises lead to overextension of our concept "ideal apartment" to "optimal apartment". This leads to the assumption that our membership judgments in a complex concept are lenient as a function of the number of constituent concepts.

Chater et al. investigated this hypothesis by conducting a series of experiments. The first experiment was a replication of Hampton's experiment (1988): subjects were asked to give membership judgments in category A, category B, and their conjunction AB on a 7-point Likert scale that ranged from -3 to +3 (using Hampton's items). Similar to Hampton, the experiment consisted of two phases with one week in between. The experiment showed a large amount of non-Boolean responses: 2.9% underextensions and 15.6% overextensions. In the second experiment, similar results were found when subjects were given only 2 response options: yes and no (3.1% vs. 12.0%). The most important experiment included the same items, testing them in conjunctions consisting of three categories (e.g. A WEAPON, A TOOL AND FARM EQUIPMENT). The experiment showed that overextensions indeed increased in the event of a triple conjunction, thus showing that membership is judged more leniently the more categories make up a complex category. These results show that categories are overextended when they are in the context of being a constituent of a complex category, thus supporting the *compensation hypothesis.*

## 2.7 SUMMARY

In this paragraph we have seen that in the past, the classical logic approach to understanding complex conjunctive concepts was challenged by the overextension effect. This effect is explained by two theories: Hampton's *composite prototype theory* and Chater et al.'s *compensation hypothesis*. In Hampton's theory category membership in conjunctive concepts is based on relatedness to a composite concept. In Chater et al.'s hypothesis the overextension effect is a compensation strategy that allows us to compromise on a criterion when looking for a best fit to a list of features. The longer the list, the more likely we are to compensate. The current experiment is part of a larger project that will investigate the nature of the overextension effect further. In this pretest we will investigate the overextension effect of a newly constructed list of items and category sets using the truth-value judgment task. The next paragraph will elaborate on the method that is used for this experiment.

## 3. METHOD

This experiment investigates which items would generate the most overextension in Dutch and can thus be used in a larger scale study investigating the nature of overextension. To this end, 68 items were judged on their membership of two categories separately and for the conjunction of these two categories. Additionally, we test if and to what extent the overextension effect is affected by the order in which the participant carries out the two tasks of judging membership of categories C and U individually and the conjunction C^U.

### 3.1 DESIGN AND MATERIALS

In the current experiment we will take a look at membership judgments using the truth value judgment task. This task includes only two response options: yes and no. The TVJ task is a widely accepted experimental method that is mostly used in language acquisition research (mostly used to assess linguistic competence; Gordon, 1998). Storms et al. (1998) have shown that membership and typicality ratings correlate highly ($r$ = .95 or higher) for their dataset (cf. Hampton; 1997), allowing us to compare the current membership judgments to Hamptons' typicality ratings without the conversing our data.

For the experiment, twelve 'sets' are used. Each 'set' consists of two categories (category C and category U). For four of these 'sets', five items are used. For the other eight 'sets', six items are used, thus resulting 68 target items in total. These objects are designed in such a way that their expected membership of category C is high while being a marginal member of category U (C stands for *certain*, U for *uncertain*). For example, when we test an item in the set FOOD-PLANT, the former is the certain category in this set to which we expect predominantly yes-responses while the latter category in the set is the uncertain one to which we expect mixed responses. PEANUT, GHERKIN and RAISIN are examples of items that were tested in this set.

68 target items are included in the experiment, each of which is judged three times: as a member of category C, as a member of category U and as a member of the conjunction of both categories C^U. This means that a total of 204 judgments is made; 68 of these 204 judgments are conjunctions, while 136 are single category judgments. A preview of the target items (n=4) showed that on average 60% of the judgments would yield yes answers, while the rest would yield no-answers. These preview results gave some indication of the expected

judgments for the actual experiment, allowing us to insert the right amount (and kind) of fillers.

To get an even distribution of yes and no answers in the experiment, 24 filler items were included (divided over the twelve 'sets'). These filler items were constructed in order to generate more no answers. Thus, for fillers, the *certain* category refers to certain no responses instead of certain yes responses (as is the case for targets). For example, for the filler CHAMOMILE in the set ANIMAL-FOOD, predominantly no answers are expected for category C (ANIMAL) while mixed responses are expected for category U (FOOD). The 24 filler items lead to another 72 extra judgments (namely cat C, cat U and C^U) to which 16% yes answers were expected. This means a total of 276 sentences were included in the experiment, in an average expected division of about 50/50 percent between expected yes and no answers. A list of all target and filler items is provided in the appendix, showing their sets and CU/UC distribution in colors.

### 3.2 PROCEDURE

Each of the participants judged the membership of each of the sixty-eight target items thrice: one task involved judging its membership in both categories of the set separately, while the other task tackled membership judgment of the object in the conjunction of the two categories. These two assignments were treated as distinct tasks, labeled 'Task ONE' and 'Task TWO' respectively. Both tasks were performed in one sitting, with a short self-paced break in between. The participants were divided into four groups: half of the participants, in groups X and Z, made the judgments on the 136 items of task ONE before the 68 items of task TWO. For the other half of the participants, in groups Y and W, the order between the tasks was exactly opposite (see Table 1 below).

| | **Task ONE > Task TWO** | **Task TWO > Task ONE** |
|---|---|---|
| **Version 1** | Participant group X | Participant group Y |
| **Version 2** | Participant group Z | Participant group W |

Table 1: Participant groups in the experiment.

There are two possibilities regarding the order of the constituent categories: CU or UC. Being marked 'UC' means that in Task ONE, the item was first tested for category U (*Is X a U?*) and then for category C (*Is X a C?*). In Task TWO, it is tested for the conjunction in the same order; U^C (*Is X a U which is also a C*?). An item that was marked 'CU' was tested in the opposite order so that category C preceded category U in both tasks.

There were two versions of the experiment. For both versions, one half of the items was marked CU while the other half was marked UC. The order of constituents differed between versions: the items that were CU in version 1 are UC in version 2 and vice versa. As an illustration of how this works, I will discuss the item *raisin,* which is tested in the set *food* (category C) and *plant* (category U). In version 1, the item is marked CU. Thus, the participant will first be confronted with the question *Is a raisin food?*, later with *Is a raisin a plant?* and finally with *Is a raisin food that is also a plant?* This is the order in which participant groups X and Y make their judgments. In version 2, the item is marked the opposite: UC. This means that the participant will first receive the question *Is a raisin a plant?*, then with *Is a raisin food?* and finally with *Is a raisin a plant that is also food?* Participant groups Z and W judge the item in this order.

Importantly, the order of the items was fixed and constant between the two tasks. The order was pseudo-randomized in the sense that no two items in the same set came directly after each other. Also, within Task ONE, all items were judged for their first category. Then, in the same order, all items were judged for their second category. This was done to minimize interaction between judgments of category C and U when judged separately (which is possible when these judgments are close to each other in terms of time).

The experiment took place at the UiL-OTS lab to eliminate surroundings confounds. Participants completed a ZEP experiment on a computer in a phonetic experiment booth. They were instructed to give their truth value judgments to questions regarding object-category membership (e.g. *Is a raisin a plant?*). Participants were explicitly instructed to follow their intuition when answering the questions to deter participants from reflecting on their judgments and possible inconsistencies. After the instructions, the participants had the opportunity to get acquainted with the task by completing three practice items. Responses were recorded in key strokes (left shift key = no, right shift key = yes).

### 3.3 PARTICIPANTS

A total of 63 participants took part in the experiment, of which 11 were male (average age 23). Recruitment took place via the UiL-OTS participant database. All participants were non-dyslectic, right handed and student or recent graduate of Utrecht University. Participants received €5,- as a reward for their participation.

## 4. RESULTS

In the current paragraph we will present the results of our experiment in terms of absolute and relative frequencies of response triples (inconsistent and consistent responses). Also, we will take a more detailed look at the amount of over- and underextension as a function of category set and item, allowing us to present a taxonomy of items.

### 4.1 RELIABILITY

A Spearman-Brown split-half reliability test was chosen to investigate how consistent the results on our experiment were; the outcome coefficient tells us how much of the test results are due to poor test construction. The target variables were divided into two halves, so that the correlation between these two random halves could be calculated. The test resulted in a reliability coefficient of .871 (equal length, between all 68 target variables), indicating adequate reliability. A Cronbach's alpha test showed a similar result ($\alpha$=.890), indicating very good internal consistency.

### 4.2 OUTLIERS AND COMMENTS

Outliers are defined as participants that differ significantly from the average frequencies of normal responses, overextensions and underextensions. To identify these outliers, the difference between the participants' individual score and the overall average (in terms of amounts of normal and over- and underextension responses) was calculated. To determine if a participant differed significantly from the average, a chi square test was performed. 18 outliers were deleted based on their chi square outcomes ($p<.01$), so that 45 participants were taken into account in the further analysis. Unfortunately, this means that 29 percent of the participants are marked as outliers. The discussion section will elaborate further on the consequences of this outlier selection.

After having finished the experiment, participants were asked if they had any comments. 9 participants mentioned that they were confused by the category *instrument*. In Dutch this word can get a wide or a narrow interpretation: when it is used in the wide sense, it may include all kinds of tools that can be used in a wide range of actions. The narrow sense of the word includes only musical instruments. The latter meaning was the intended when

designing the experiment. The fact that the *instrument* can have these two different meanings did not appear when previewing the experiment (n=4).

### 4.3 DATA CODING

The dependent variable in this experiment is the judgment of the participant regarding a statement on item membership in two simple categories and one complex category (which is a conjunction of the latter two). For each item, three dichotomous decisions are made so that $2^3$=8 possible response patterns emerge. According to classical set conjunction, we may only agree with membership in the set iff we agree with membership in both constituent categories. If either one (or both) of the constituents does not receive positive membership judgment, we must reject membership in the conjunction (see again Figure 1 for the truth table for conjunctions). We can label the eight possible response patterns as *consistent* or *inconsistent* with this logic. The *consistent* responses are [---], [+++], [-+-] and [+--]. The *inconsistent* responses are [--+], [++-], [-++] and [+-+].

| **Response double [cat C, cat U]** | **Response triple [cat C, cat U, conj]** | Absolute frequency | Relative frequency in % of all responses |
|---|---|---|---|
| [++] | [+++][a] | 1037 | 33,9 |
| | [++-][b1] | 129 | 4,2 |
| [--] | [---][a] | 219 | 7,2 |
| | [--+][b2] | 25 | ,8 |
| [-+] | [-+-][a] | 108 | 3,5 |
| | [-++][b2] | 43 | 1,4 |
| [+-] | [+--][a] | 1152 | 37,6 |
| | [+-+][b2] | 347 | 11,3 |

Table 2: Frequencies of each type of response triple across all items and subjects [C, U, conjunction].
[a]Consistent. [b]Inconsistent. [1]Underextension. [2]Overextension.

### 4.4 Frequencies of response triples

The frequencies of all eight response patterns are summarized in Table 2. The data show that consistent [+--] and [+++] are the most frequent patterns overall. [+-+] is by far the most frequent inconsistent pattern (11.3%).

Furthermore, Table 2 shows that the consistent variant of the triple is always more frequent than the inconsistent variant. For example, the consistent [+--] pattern forms 37.6% of all responses while only 11.3% of all responses are of the inconsistent alternative form [+-+]. We can cluster the response triples into three distinct classes: normal responses, overextension responses and underextension responses. Table 3 shows the clustered frequencies for these three categories across all items and participants.

| **Response class** | **Absolute frequency** | **Relative frequency (% of all responses)** |
|---|---|---|
| Normal | 2516 | 82.2 |
| Overextension | 415 | 13.6 |
| Underextension | 129 | 4.2 |
| Total | 3060 | 100 |

Table 3: Clustered frequencies of each class of response triples across all items and subjects. Normal includes [+++], [---], [+--] and [-+-]. Overextension includes [--+], [+-+] and [-++]. Underextension includes [++-].

As we can see in Table 3, the vast majority of responses (82.2%) are normal, and thus consistent with Boolean logic. We see that 17.8% of all responses are inconsistent with Boolean logic: these inconsistent responses are primarily overextensions (13.6%) while only few are underextensions (4.2%).

### 4.5 Inconsistent responses per category set

Figure 2 shows the number of over- and underextensions as a function of the category set. We see that the number of overextensions is highest for the category sets food-plant, tool-weapon, weapon-tool, kitchen utensil-instrument and organ-food. The category set liquid-beverage shows low frequencies of overextensions. Underextensions are most frequent in the categories clothing-footwear , liquid-beverage and kitchen utensil-instrument.

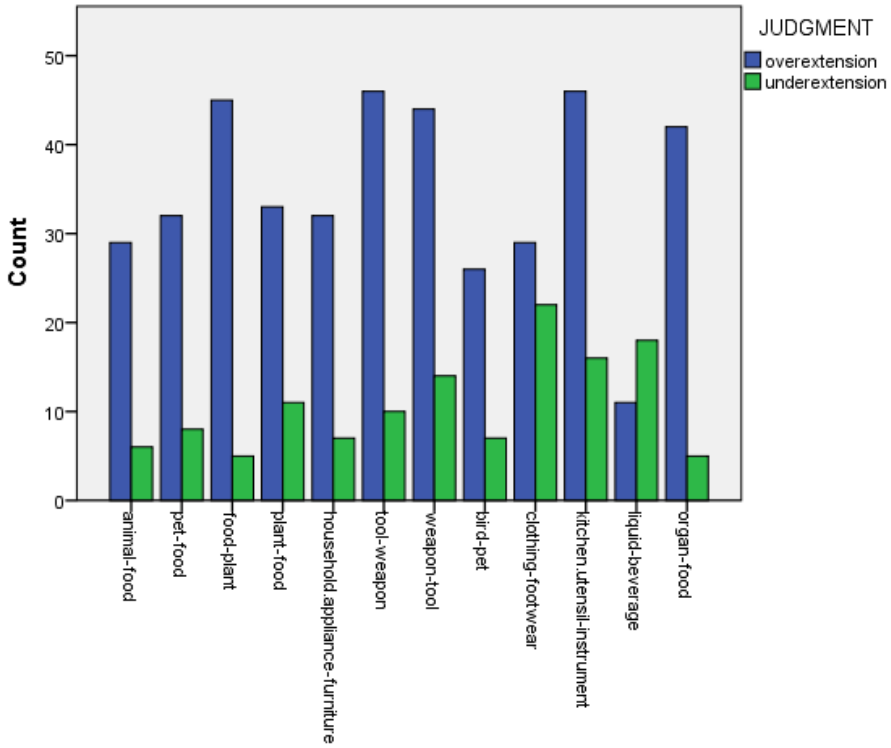


Figure 2: Absolute frequencies of overextension and underextension responses as a function of category set.

### 4.6 INCONSISTENT RESPONSES PER ITEM

For 13 out of 68 items overextension patterns are the only non-logical responses (ALOËVERA, BAMBOO, BROTH, CAT, DOG, GINGER, HIFI, NETTLE, RAVEN, SAUERKRAUT, THYMUS, TONGUE AND WASHING MACHINE) while 1 item generates exclusively underextension as a non-logical response (ALCOHOL).

We can take a more extensive look at the frequencies of over- and underextensions per item. A table containing frequencies for all items is included in Appendix A. The distribution of overextension responses per item is visualized in Figure 3 below.

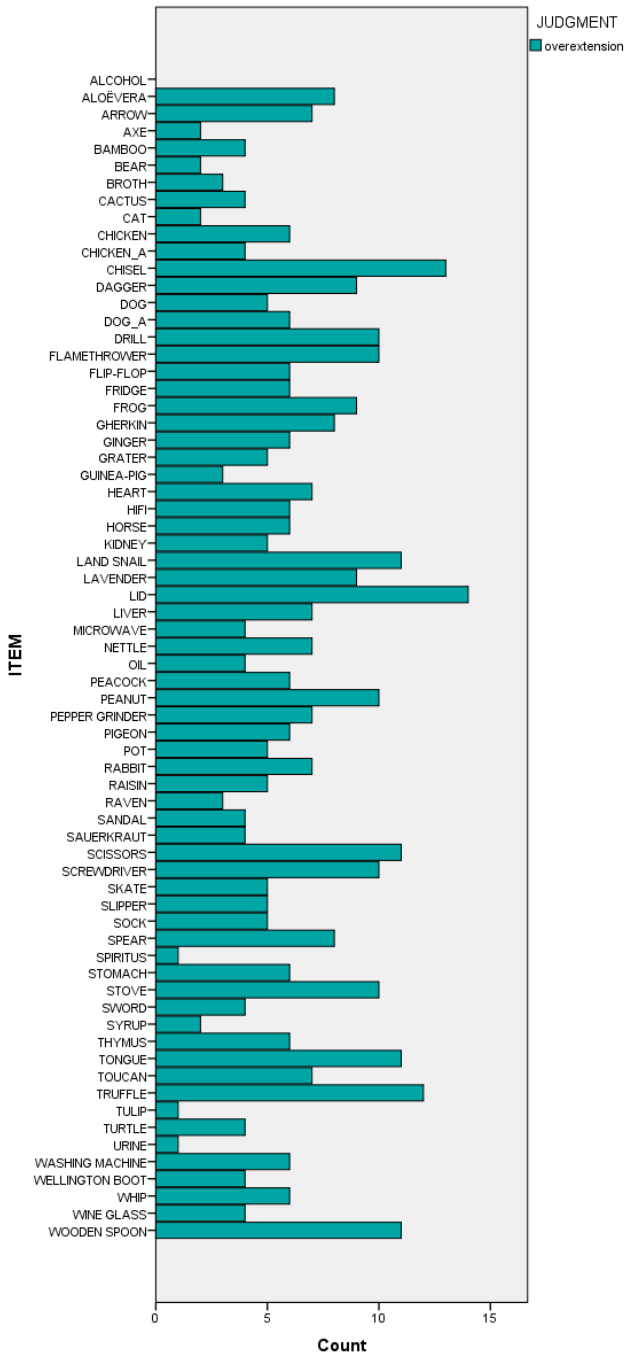


Figure 3: Frequency of overextension responses per item as a function of item.

On average, 13.6% of the responses are overextensions. Twenty-five items show a higher amount of overextension responses: 15.6% (ARROW, HEART, LIVER, NETTLE, PEPPER GRINDER, RABBIT, TOUCAN), 17.8% (ALOËVERA, GHERKIN, SPEAR), 20% (DAGGER, FROG, LAVENDER), 22% (DRILL, FLAMETHROWER, PEANUT, SCREWDRIVER, STOVE), 24% (LAND SNAIL, SCISSORS, TONGUE, WOODEN SPOON), 26.7% (WOODEN SPOON), 28.9% (CHISEL) and even 31.1% (LID).

Underextension patterns emerge on average for 4.4% of the responses. Frequencies of underextensions per item are visualized in Figure 4. Twenty-two items generated more underextensions: 6.7% (TULIP, AXE, MICROWAVE, OIL, FLIP-FLOP, FRIDGE, HORSE, PIGEON, SPEAR, DRILL), 8.8% (SYRUP, CACTUS, WELLINGTON BOOT, GRATER, POT, WHIP, PEPPER GRINDER, LAVENDER), 11.1% (SPIRITUS, URINE) and 13.3% (SLIPPER, SOCK).

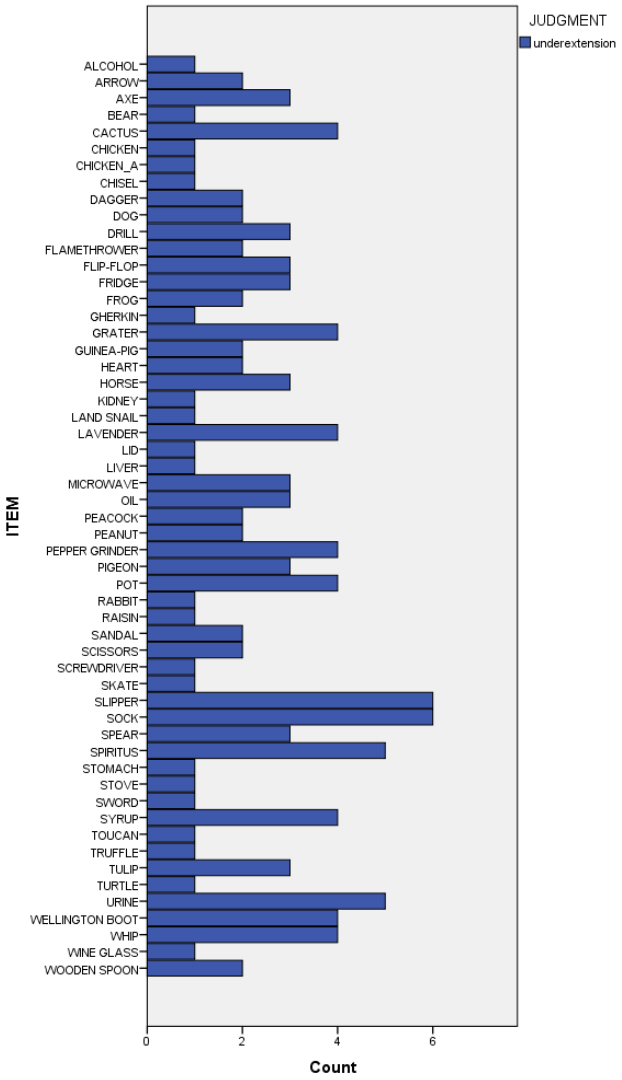


Figure 4: Frequency of underextension responses per item. Items cat, broth, raven, bamboo, sauerkraut, dog_A, ginger, hifi, thymus, washing machine, nettle, aloëvera and tongue are not included in the graph, since these items showed no underextension responses at all.

## 4.7 TAXONOMY OF ITEMS

Using the data that is summarized in Figure 3 and Figure 4, we can construct a taxonomy of items. This taxonomy classifies items as ‘sure’ or ‘unsure’ items; the first category is marked ‘normal’ items and shows predominantly normal responses. The second other category consists of vaguer items, to which participants respond at or above average levels for over-

and underextension (13.6% and 4.2% respectively, see Table 3). These vaguer items will be further divided into three subcategories: *overextension items* (O) that score high on overextension, *underextension items* (U) that score high on underextension and *abnormal items* (A) that score high on both over- and underextension. The threshold for a high score is put at 75% of the average for that category (10% of all responses for label 'overextension' and 3% of all responses for label 'underextension'). The distribution of all items in these categories is shown in Table 4.

| | **Category (# of items)** | **Items** |
|---|---|---|
| Sure items | Normal items (11) | alcohol, bamboo, bear, broth, cat, chicken_A, raven, sauerkraut, sword, turtle, wine glass |
| Unsure items | Overextension items (23) | aloëvera, chicken, chisel, dog_A, gherkin, ginger, hifi, kidney, landsnail, lid, liver, nettle, rabbit, raisin, screwdriver, skate, stomach, stove, thymus, tongue, toucan, truffle, washing machine |
| | Underextension items (11) | axe, cactus, guinea-pig, microwave, oil, sandal, spiritus, syrup, tulip, urine, wellington boot |
| | Abnormal items (23) | arrow, dagger, dog, drill, flamethrower, flip-flop, fridge, frog, grater, heart, horse, lavender, peacock, peanut, pepper grinder, pigeon, pot, scissors, slipper, sock, spear, whip, wooden spoon |

Table 4: Items distributed over the categories N, O, U and A.

### 4.8 Order and version effects

A one-way ANOVA ($F(1,43) = 7.012$, $p = .011$) showed that the number of overextensions differs significantly per order: there are significantly more overextensions in the separate > conjunction order than in the conjunction > separate order. There is no effect of order on the amount of the normal ($F(1,43) = 4.074$, $p = .050$) and underextension ($F(1,43) = .990$, $p = .325$) responses. Another ANOVA test revealed that there was no significant effect of version

on the amount of normal ($F(1,43) = .042$, $p=.838$), overextension ($F(1,43) = .004$, $p = .951$) or underextension responses ($F(1,43) = .353$, $p = .556$).

### 4.9 SUMMARY

This section has shown the experimental results in terms of relative and absolute frequencies of different kinds of response triples and response classes. We have seen that there was a high amount of intra subject variability, which was reflected in the high amount of outliers for our analysis. 17.8% of all responses were inconsistent with classical logic, of which 13.6% were overextensions and 4.2% were underextensions. A taxonomy of items was presented, showing that only 11 of 68 items showed low amounts of over- and underextensions, while 23 items showed high amounts of inconsistent in both categories. In addition, we have seen that there was an effect of order, but no effect of version on the amount of over- and underextensions. The following section will take a more detailed look at the data, including a discussion on the assumptions (models) underlying the analysis.

## 5. ANALYSIS AND DISCUSSION

In the previous paragraph we have seen that 17.8% of all responses were inconsistent with classical logic: 13.6% overextension and 4.2% underextension. In this paragraph we will investigate whether there is a significant difference between the number of over- and underextensions in the current dataset. Such a significant difference would prove that we have indeed found the overextension effect, thus replicating Hampton's results. In the first two sections we will provide an analytical background. The third and fourth section include analyses of the data under the two models that are proposed.

### 5.1 ANALYTICAL BACKGROUND: M1

Hampton (1988) and Chater et al. (1990) investigated to what extent their data differed from classical conjunction. Hampton derived the prediction that unreliability in responding while applying a conjunctive rule should create equal frequencies of over- and underextension. This unreliability would be due to a random change of mind in the participant when judging membership in the complex category as compared to the constituent category judgments (we will call this hypothesis M1). For example, a participant can agree with RAISIN in the complex category FOOD THAT IS ALSO A PLANT while at the same time negating its membership in the constituent category PLANT. If such a deviation from the classical model is seen as a random change of mind in one (or both) of the categories, equal frequencies of over- and underextension are expected.

These expected equal frequencies are illustrated as follows. For any item there are 4 possible response patterns for their judgments on the constituent categories C and U: [++], [+-], [-+] and [--]. When participants perform the second judgment task, they may change their mind about just C, just U or about both. All possible options are shown in Table 5 below.

| | **Change mind about C** | **Change mind about U** | **Change mind about both** |
|---|---|---|---|
| [++] | [++] □ [-+] □ [-]<br>[++-] under | [++] □ [+-] □ [-]<br>[++-] under | [++] □ [--] □ [-]<br>[++-] under |
| [+-] | [+-] □ [--] □ [-]<br>[+--] normal | [+-] □ [++] □ [+]<br>[+-+] over | [+-] □ [-+] □ [-]<br>[+--] normal |
| [-+] | [-+] □ [++] □ [+] | [-+] □ [--] □ [-] | [-+] □ [+-] □ [-] |

| | | | |
|---|---|---|---|
| | [-++] over | [-+-] normal | [-+-] normal |
| [--] | [--] □ [+-] □ [-]<br>[---] normal | [--] □ [-+] □ [-]<br>[---] normal | [--] □ [++] □ [+]<br>[--+] over |

Table 5: Results of a random change in mind for category C, category U or both; as theorized by Hampton (1988). The leftmost column contains the initial response double to category C (1st token) and category U (2nd token). Each cell contains: [initial response double] > [response double with change-of-mind] > [response to the conjunction, following classical logic]. The blue text shows the conclusive response triple that combines the original response double and the newly calculated conjunction response, accompanied with a label (normal/overextension/underextension).

In Table 5 we see that following this assumption, we would expect similar frequencies of over- and underextensions in the dataset: half of the responses show normal patterns, while overextensions and underextensions both form 25 percent of the responses. This similarity allows us to use our unweighted over- and underextension frequencies for the further analysis. Following this model, an uneven distribution between over- and underextensions in a dataset would yield that people are prone to change their minds in a specific direction, indicating an effect.

### 5.2 THE ALTERNATIVE: M2

As an alternative to the model described above, there is a possibility that unreliability in responding (and in applying a conjunctive rule) should create unequal frequencies of over- and underextensions. After having completed the current experiment, the majority of participants reported having difficulties giving their judgments. Participants were explicitly instructed not to overthink their answers and to follow their intuition when responding to the statements. Some participants reported being inclined to answer as quickly as possible, which caused them to enter unintended responses. Such responses are definitely arbitrary and can be seen as a reflex rather than reflecting a membership judgment. It is a possibility that the fact that participants deviate from classical logic is due to such factors. Therefore, their behavior can be considered arbitrary, reflecting noise rather than bona fide judgments. If this were the case, we would expect the frequencies of normal, overextension and underextension responses to be at chance level, as is illustrated in Table 6.

| **Response triple** | **Label and chance level** |
|---|---|
| [---] | Normal (50%) |
| [+++] | |
| [-+-] | |
| [+--] | |
| [-++] | Overextension (37.5%) |
| [+-+] | |
| [--+] | |
| [++-] | Underextension (12.5%) |

Table 6: Probabilities of arriving at normal, overextension and underextension responses when guessing on all three judgments [C, U, conj].

Table 6 shows the chances of reaching a certain response pattern when guessing for all three judgments [C, U, conjunction]. For example, if a participant would put in arbitrary responses on the three judgments of an item in the categories FOOD, PLANT and FOOD THAT IS ALSO A PLANT, he would have 12.5% chance to show an underextension response pattern while having a 37.5% chance to arrive at an overextension response triple.

For the reasons mentioned above, it does not make sense to use unweighted frequencies in order to check if overextension appears in a significantly different pattern than underextension in our dataset. Under M2, we can only draw conclusions if we weigh the frequencies of normal, overextension and underextension responses according to the chance proportions in Table 6. This means dividing the normal frequencies by 4 and dividing overextension frequencies by 3. Weighted frequencies of the data are shown in Table 7 below.

| | UNWEIGHTED | | WEIGHTED | |
|---|---|---|---|---|
| | **Absolute frequency** | **Relative frequency (% of all responses)** | **Absolute frequency** | **Relative frequency (% of all responses)** |
| Normal[a] | 2516 | 82.2% | 629 | 70.2% |
| Overextension[b] | 415 | 13.6% | 138 | 15.4% |
| Underextension[c] | 129 | 4.2% | 129 | 14.4% |
| Total | 3060 | 100% | 896 | 100% |

Table 7: Unweighted and weighted frequencies of normal, overextension and underextension responses. Weighted frequencies are derived by dividing unweighted frequencies by [a]4, [b]3 and [c]1.

If we take a brief look at the data in Table 7, we already see that the weighted frequencies show a completely different distribution of overextension and underextension patterns than unweighted frequencies: 15.4% and 14.4% compared to 13.6% and 4.2% respectively. Further analysis of the data will show what are the implications of the assumption of M2 for our results.

### 5.3 Analysis of the data under M1

To analyze the data, a Wilcoxon signed-rank test was chosen. This test was selected because the amount of over- and underextensions in our data are not entirely independent: a participant cannot both over- and underextend one item. The data used are frequencies of the response categories normal, overextension and underextension, so that a nonparametric test must be used (see Chater et al., 1990:499), in this case the Wilcoxon-signed rank test.

Under M1, 4.2% of the responses were underextensions while 13.6% of the responses were overextensions (summed over all subjects). The Wilcoxon signed-rank test was applied to the frequency data across all items, showing a significant difference between overextensions and underextensions (Z=-5.385, N=45, p=.000), with a large effect size (r=-.56). This test shows that overextension is a far more common pattern than underextension in our unweighted dataset. Thus, when we assume M1, the overextension effect is present in our

dataset. Hampton's (1988) finding that people do not follow Boolean-logic when judging conjunctive complex category memberships is replicated, indicating the overextension effect.

A second analysis was performed in the form of a chi square test of goodness-of-fit on the frequency data, comparing observed and expected count for the overextension and underextension patterns. Expected count in this test is taken to be an even distribution of over- and underextensions, following the M1 hypothesis. The chi square test of goodness-of-fit showed that over- and underextension were not evenly distributed in our data $X^2(2, N=3060) =150.174$, $p<.0001$. This analysis again shows that the overextension effect is present in our *unweighted* dataset.

### 5.4 Analysis of the data under M2

For the re-analysis, the data are weighted as in Table 7. Now, summed over all subjects, 14.4% of the responses were underextensions while 15.4% of the responses were overextensions. The same Wilcoxon signed-rank test was applied to the weighted frequency data across all items, showing no significant difference between overextensions and underextensions ($Z=-.742$, $N=45$, $p=.458$). Along the lines of M2, overextension is not different from underextension.

The weighted dataset was subjected to a similar chi square test of goodness-of-fit in order to investigate if the overextensions and underextensions were evenly distributed in our data. Again, expected counts were based on an even distribution of over- and underextensions. The chi square test showed that the data were not significantly different from an even distribution of over- and underextensions $X^2(2, N=896) =.303$, $p=.8593$. Therefore, our *weighted* dataset again does not support the idea of an overextension effect.

### 5.5 Alternative outlier selection

In the results section we have seen that 29 percent of the participants were marked as outliers when the definition included participants that differed significantly from the average amounts of over- and underextension. Another definition of outliers was applied to the dataset to check if this would lead to different outcomes. The new definition included participants that showed a relatively low amount of normal responses to the filler items. Filler items were judged the categories C (clear non-member) and U (unclear marginal member). Inconsistent responses to such filler items indicate that participants changed their judgment about the clear non-

membership of the filler in category C, which should not be possible (regarding the nature of the items). Such responses indicate the participant is not paying attention or does not understand the task. Participants that were not able to respond with at least 80% normal responses on the fillers were deleted (n=20) for the alternative analysis, such that n=43. A replication of the analysis in sections 5.3 and 5.4 using the remaining subjects did not lead to different outcomes.

### 5.6 Intermediate conclusion

As we have seen in this section, there are two possible analyses of our data: one using raw frequency data and another using weighted frequency data. Importantly, we have seen that these analyses lead to different outcomes regarding the difference between over- and underextension. If we follow the M1 (change-of-mind) hypothesis, we see that overextension appears in our data as a significant effect, while underextension does not. However, if we follow the M2 (noise) hypothesis, both over- and underextension appear at chance level in our data. In other words, the assumptions that underlie our analysis have great impact on the outcome.

## 6. FURTHER THEORETICAL QUESTIONS

In the previous section, we have seen that overextension comes out as a significant effect in our data when we apply a statistical analysis based on M1, while underextension does not. When we apply M2, overextension and underextension come out as chance-level effects within our dataset. How can we understand these results in the light of theories of categorization and complex concepts?

### 6.1 BETWEEN SUBJECTS VARIATION

In our dataset, we have seen variation between subjects, which is illustrated by the high number of outliers (29%, based on their difference from the average scores) in the data. In the introduction we have seen that Hampton's (1988) Threshold Model can be used to account for such variation in category membership judgments. The model includes a prototype model of concepts, measurement of similarity and a threshold for category membership. Following the threshold model, variance in categorizations may appear between subjects coming from three aspects of the process of categorization: the representation of the instance *x*, the representation of the composite concept *A* or the threshold that is applied for membership to the complex category (Barsalou, 1987; Hampton, 1995).

In his paper on the instability of graded structures, Barsalou (1987) discussed a meta-analysis, investigating between subject agreement in typicality ratings in experiments on simple concepts. His analysis showed that participants show, on average, only .50 correlation with another participant's typicality ratings. Barsalou concludes from these findings that typicality is highly variable across individuals. Unfortunately, such correlation values cannot be calculated for our dataset due to the fact that we are dealing with categorical data. However, we may interpret the variance in our dataset as an indication for the instability of complex categories across subjects.

### 6.2 INCONSISTENT RESPONSES UNDER M1

In the analysis section, we have seen that 17.8% of participants' responses are inconsistent with classical conjunction under M1. Such inconsistent responses are hypothesized both by M1 and M2. In M1, we assume participants change their minds about the membership judgments of either of the constituent categories. In the past, Hampton (1988) investigated to

what extent people show inconsistent responses when judging item membership in two categories and their conjunction. His experiments have shown that the majority of the violations of the classical conjunction rule were overextensions, while finding only few underextensions. Driven by the uneven distribution of over- and underextensions in the dataset, Hampton concluded that his results did not indicate simple changes of minds. Namely, if the inconsistent responses would have been due to random changes of mind, we should have seen an even distribution of over- and underextensions. Rather, following Hampton's line of reasoning, the high amount of overextensions in his dataset can be attributed to the conceptual representation that is formed for the complex concept.

In Hampton's Composite Prototype account, this conceptual representation is a merger of the attributes of the two constituent concepts, forming a novel composite prototype that represents the conjunction as a whole. This account makes use of the *intensions* of constituent concepts ("sets of generic properties characterizing the kind in question"; Hampton, 2013:2). Because the composite prototype differs (to some extent) from both constituents, items can be overextended. Under M1, our analyses show that participants are indeed prone to overextend their categories, supporting Hampton's findings. The fact that under M2, previous findings are not supported by our data, will be discussed further in the next paragraph.

### 6.2 Inconsistent responses under M2

Similar to M1, M2 hypothesizes inconsistent responses. In contrast with the analysis under M1, M2 hypothesizes that these inconsistencies do not show differing patterns. Analysis of the weighted dataset has shown there is indeed an equal distribution of over- and underextensions, both appearing at chance-level. Therefore, participants' responses do appear to reflect random patterns, indicating overextension (and underextension) are the result of other factors than people's bona fide judgments. Thus, Hampton's composite prototype model cannot account for these data.

The relative frequencies of over- and underextensions that are hypothesized by M2, can be found in other datasets than the current one. The experiments by Hampton (1988) and Chater et al. (1990) show similar distribution of over- and underextensions: overextensions appear roughly three times more often than underextensions (reflecting the assumptions of M2, see Table 6). It would be interesting to re-analyze the data of these experiments, in order to see if we arrive at similar results.

Chater at al. (1990) consider the inconsistent responses to be caused by a best fit strategy, allowing for membership in the conjunction in a more loose fashion than when we judge membership in the constituent category separately. In light of this hypothesis, conceptual combination is considered to be a context effect: each constituent of the complex category has to be evaluated in the context of the other.

### 6.3 Other explanations

Because neither Hamptons' nor Chater et al.'s theories can deal with the outcome of the analysis under M2, we may take a look at other explanations for the inconsistent responses. High within subject variability has been found in other experiments investigating typicality and category membership.

Barsalou and Medin (1986) investigated to what extent participants' typicality ratings were consistent across sessions for simple categories. In an experiment, participants were asked to give typicality judgments for items in categories at two points in time (pause ranging from two hours to four weeks). The same items were judged in the same categories in both tasks. The experiment showed that intra subject reliability decreased over time. Moreover, the variability of the rating was highest for moderately typical items (as compared to low variability for highly typical or atypical items). Following these findings, we can hypothesize that the inconsistent responses in the current experiment may have been triggered by the fact that all items were moderately typical for at least one constituent category (category U).

Similarly, McCloskey and Glucksberg (1978) have shown that people use different information for determining membership on every judgment occasion. They asked participants to give their membership judgments on 540 category-name pairs, including highly typical category members (*chair-furniture*), unrelated items (*cucumber-furniture*) and items that were intermediate typical (*bookend-furniture*). Their experiment showed intra- and inter subject agreement was high for clear cut cases (high typical and unrelated items), while intermediate items showed great variability. Again, the fact that all items were moderately typical for at least one constituent category (category U) may have triggered inconsistent responses.

Both Barsalou's and McCloskey and Glucksbergs findings indicate that intra subject variability is common in categorization tasks. Similarly, the current experiment consisted almost entirely of moderately typical items (since each item was judged in an uncertain

category), which hay have led to confusion in participants, giving rise to guess patterns in our dataset.

### 6.4 SUGGESTED FUTURE RESEARCH

Initially, the current experiment was set up as a pretest in a larger-scale project. The aim of this project was to investigate the nature of the overextension phenomenon. As we have seen in the previous sections, analyses under M1 and M2 demonstrate different outcomes. The problem we are confronted with following this finding is: how do we know which of the models is correct? From the current experiment we are not able to choose whether we should look at overextension as the result of a strategy for dealing with complex concepts or as the result of noise. Further research should investigate this (i.e. the applicability of M1 vs. M2) before moving forward in examining the nature of the overextension effect.

Only if and after we can prove that overextension is not merely noise, we can further investigate the two hypotheses regarding the nature of the effect: either overextension is a compositional mechanism that is triggered by our computation of a composite prototype for complex concept (cf. Hampton, 1997) or it may be a more general *compensation* strategy that is triggered by the complexness of the task that people perform (cf. Chater et al., 1990).

## 7. CONCLUSION

This thesis was aimed at replicating Hampton's experiment, investigating the overextension effect for Dutch using a truth-value judgment task. We have seen that consistent responses were by far the most common in our experiment (82.2%), while some were overextensions (13.6%) and few were underextensions (4.2%). The most common consistent response pattern in our dataset was [+--] (37.6%), while the most common inconsistent response triple was the overextension [+-+] (11.3%). Analyses of the data similar to Hampton (1988) and Chater et al. (1990) has shown that the overextension effect was present in the dataset: participants were prone to overextend their categories when judging membership in conjunctive complex categories. In this analysis, inconsistent responses are hypothesized to be caused by people changing their mind about either one (or both) of the constituent categories in the second part of the experiment (M1).

In previous research, the frequencies of response patterns have been taken as an indication that overextension is an effect: analyses on response frequencies have shown that overextension is significantly more frequent than underextension. However, we believe that we need to take into account that people may be operating in a non-systematic fashion, that is to say that inconsistent responses are due to noise (. Therefore, we need to consider chance levels for all response patterns when analyzing our dataset. A re-analysis was performed using weighted data. Remarkably, the difference between over- and underextension is not significant when overextension and underextension frequencies are weighted. This analysis is based on the alternative assumption that deviations from classical logic (for membership in conjunctive categories) are generally just noise (M2).

Based on the results of the current experiment, we are not able to choose between the two models for inconsistent responses. It is a is a possibility that in fact, the explanation of inconsistent responses lies in a combination of the two: some may be the result of a change-of-mind (for example, caused by a compensation effect), while others are caused by noise. The research findings summarized above show the need to investigate the nature of participants' inconsistent responses in more detail before exploring the overextension effect further.

## 8. REFERENCES


Barsalou, L. W. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 101-140). New York: Cambridge University Press.

Barsalou, L. W., & Medin, D. L. (1986). Concepts: Static definitions or context-dependent representations? *Cahiers De Psychologie, 6*, 187-202.

Chater, N., Lyon, K., & Myers, T. (1990). Why are conjunctive categories overextended? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(3), 497-508.

Gordon, P. (1998). The truth-value judgment task. Paper presented at the *Methods for Assessing Children's Syntax,*

Hampton, J. A. (2103). Conceptual combination: Extension and intension. commentary on Aerts, Gabora, and Sozzo. *Topics in Cognitive Science, 4*(1), 1-5.

Hampton, J. A. (1995). Testing the prototype theory of concepts. *Journal of Memory and Language, 34*(5), 686-708.

Hampton, J. A. (1997). Conceptual combination: Conjunction and negation of natural concepts. *Memory & Cognition, 25*(6), 888-909.

Hampton, J. A. (1988). Overextension of conjunctive concepts: Evidence for a unitary model of concept typicality and class inclusion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(1), 12-32.

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science, 31*(3), 355-384.

Łukasiewicz, J. (1968). On three-valued logic. *The Polish Review,* 43-44.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition, 6*(4), 462-472.

Smith, E. E., & Osherson, D. N. (1984). Conceptual combination with prototype concepts. *Cognitive Science, 8*(4), 337-361.

Smith, E. E., Osherson, D. N., Rips, L. J., & Keane, M. (1988). Combining prototypes: A selective modification model. *Cognitive Science, 12*(4), 485-527.

Storms, G., De Boeck, P., Van Mechelen, I., & Ruts, W. (1998). Not guppies, nor goldfish, but tumble dryers, noriega, jesse jackson, panties, car crashes, bird books, and stevie wonder. *Memory & Cognition, 26*(1), 143-145.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*(3), 338-353.

Appendix A: Absolute frequencies of the three judgments patterns (normal, overextension, underextension) per item.

| | JUDGMENT | | | Total |
|---|---|---|---|---|
| | normal | overextension | underextension | |
| ALCOHOL | 44 | 0 | 1 | 45 |
| ALOËVERA | 37 | 8 | 0 | 45 |
| ARROW | 36 | 7 | 2 | 45 |
| AXE | 40 | 2 | 3 | 45 |
| BAMBOO | 41 | 4 | 0 | 45 |
| BEAR | 42 | 2 | 1 | 45 |
| BROTH | 42 | 3 | 0 | 45 |
| CACTUS | 37 | 4 | 4 | 45 |
| CAT | 43 | 2 | 0 | 45 |
| CHICKEN | 38 | 6 | 1 | 45 |
| CHICKEN_A | 40 | 4 | 1 | 45 |
| CHISEL | 31 | 13 | 1 | 45 |
| DAGGER | 34 | 9 | 2 | 45 |
| DOG | 38 | 5 | 2 | 45 |
| DOG_A | 39 | 6 | 0 | 45 |
| DRILL | 32 | 10 | 3 | 45 |
| FLAMETHROWER | 33 | 10 | 2 | 45 |
| FLIP-FLOP | 36 | 6 | 3 | 45 |
| FRIDGE | 36 | 6 | 3 | 45 |
| FROG | 34 | 9 | 2 | 45 |
| GHERKIN | 36 | 8 | 1 | 45 |
| GINGER | 39 | 6 | 0 | 45 |
| GRATER | 36 | 5 | 4 | 45 |
| GUINEA-PIG | 40 | 3 | 2 | 45 |
| HEART | 36 | 7 | 2 | 45 |
| HIFI | 39 | 6 | 0 | 45 |
| HORSE | 36 | 6 | 3 | 45 |
| KIDNEY | 39 | 5 | 1 | 45 |

| | | | | | |
|---|---|---|---|---|---|
| | LAND SNAIL | 33 | 11 | 1 | 45 |
| | LAVENDER | 32 | 9 | 4 | 45 |
| | LID | 30 | 14 | 1 | 45 |
| | LIVER | 37 | 7 | 1 | 45 |
| | MICROWAVE | 38 | 4 | 3 | 45 |
| | NETTLE | 38 | 7 | 0 | 45 |
| | OIL | 38 | 4 | 3 | 45 |
| | PEACOCK | 37 | 6 | 2 | 45 |
| | PEANUT | 33 | 10 | 2 | 45 |
| | PEPPER GRINDER | 34 | 7 | 4 | 45 |
| | PIGEON | 36 | 6 | 3 | 45 |
| | POT | 36 | 5 | 4 | 45 |
| | RABBIT | 37 | 7 | 1 | 45 |
| | RAISIN | 39 | 5 | 1 | 45 |
| | RAVEN | 42 | 3 | 0 | 45 |
| | SANDAL | 39 | 4 | 2 | 45 |
| | SAUERKRAUT | 41 | 4 | 0 | 45 |
| | SCISSORS | 32 | 11 | 2 | 45 |
| | SCREWDRIVER | 34 | 10 | 1 | 45 |
| | SKATE | 39 | 5 | 1 | 45 |
| | SLIPPER | 34 | 5 | 6 | 45 |
| | SOCK | 34 | 5 | 6 | 45 |
| | SPEAR | 34 | 8 | 3 | 45 |
| | SPIRITUS | 39 | 1 | 5 | 45 |
| | STOMACH | 38 | 6 | 1 | 45 |
| | STOVE | 34 | 10 | 1 | 45 |
| | SWORD | 40 | 4 | 1 | 45 |
| | SYRUP | 39 | 2 | 4 | 45 |
| | THYMUS | 39 | 6 | 0 | 45 |
| | TONGUE | 34 | 11 | 0 | 45 |
| | TOUCAN | 37 | 7 | 1 | 45 |
| | TRUFFLE | 32 | 12 | 1 | 45 |
| | TULIP | 41 | 1 | 3 | 45 |
| | TURTLE | 40 | 4 | 1 | 45 |

| | | | | | |
|---|---|---|---|---|---|
| | URINE | 39 | 1 | 5 | 45 |
| | WASHING MACHINE | 39 | 6 | 0 | 45 |
| | WELLINGTON BOOT | 37 | 4 | 4 | 45 |
| | WHIP | 35 | 6 | 4 | 45 |
| | WINE GLASS | 40 | 4 | 1 | 45 |
| | WOODEN SPOON | 32 | 11 | 2 | 45 |
| Total | | 2516 | 415 | 129 | 3060 |