Institute of Philosophy
Collaborative Research Center 991
Heinrich-Heine-University Düsseldorf

Prototype frames

A probabilistic account of typicality

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Philosophie (Dr. phil.)
durch die Philosophische Fakultät der
Heinrich-Heine-Universität Düsseldorf

Vorgelegt von
Annika Schuster

aus
Essen

Betreuer:
Prof. Gerhard Schurz
Prof. James Hampton

Düsseldorf Mai 2022

Acknowledgements

This dissertation builds on the work of my supervisor, Professor Gerhard Schurz, who has employed and supported me for many years. His invaluable advice and constant mentoring made this work possible, and I am deeply grateful to him.

I also extend my heartfelt thanks to Professor James Hampton for the time he devoted to discussing and significantly improving both the empirical and theoretical parts of this investigation, as well as for his continued support.

Many thanks to Dr. Corina Strößner for her ongoing guidance, thoughtful advice, and honest feedback. She helped me immensely and was always willing to listen.

Research for this dissertation was generously supported by the German Research Foundation (DFG), grant number SFB991-D01. The exchange with members of the Collaborative Research Center SFB991 was indispensable to my work, and I thank them all.

I thank Benoît, Sophie and Hannah for both supporting my work and distracting me from it, and the home full of love and discussions we built together.

Abstract

In the 1970s, the psychologist Eleanor Rosch and her team found in a series of experiments that people overwhelmingly agree which category members are typical for their category (C) and which are not – apples are typical fruit, hammers are typical tools, robins are typical birds and chairs are typical furniture, while olives, scissors, emus and refrigerators are atypical examples of these categories. The typicality ordering of a C that is obtained from mean typicality ratings for many subcategories (SCs) is assumed to indicate an important component of the structure of mental representations of Cs, category concepts: in line with the thoughts presented in late Wittgenstein's natural language philosophy, it is commonly assumed that there are additional, typicality-contributing properties stored in category concepts. Classical definitions, which consist of singly necessary and jointly sufficient conditions, cannot explain the internal structure of category concepts, because the fulfilment of these conditions makes all members equally representative of C. Typicality-contributing properties are shared between subsets of SCs, albeit not present in all of them, and constitute a family resemblance relationship between SCs, much like the physical features shared between family members. In addition to being the ripened reproductive bodies of a seed plant, apples, like many other fruits, taste sweet and have bright colours and olives do not, and in addition to being warm-blooded egg-laying vertebrates, robins are small, sing and fly while emus are big and do not sing or fly. Furthermore, not all properties are equally important for typicality. A food item that is sweet will be rated as more typical than one that is brightly coloured and a bird that flies will be rated as more typical than one that is small. A formal model of category representation that contains typical properties and reflects the extent of their typicality-contribution provides an enriched picture of mental representations and is a suitable basis for the investigation of phenomena like conceptual combination and reasoning patterns involving category concepts.

In this thesis, I develop such a model: probabilistic prototype frames. I investigate theoretically and empirically how typicality-contributing properties of Cs can be identified and how the extent of their typicality contribution can be quantified in a way that predicts and explains the typicality ordering of Cs. Schurz' (2001, 2005, 2012) evolution-theoretic account of normality offers an intriguing way to identify typicality-contributing properties and base reasoning with prototype concepts on an ontologically justified probabilistic framework. He argues that not only the formation of biological categories, but also that of many other common-sense categories, is guided by the evolutionary principles of selection, variation and reproduction which lead to a prototypical norm state in which most category members are most of the time. This state is characterised by prototypical properties in the wide sense (iws), for which the conditional probability of the property P given the category C, $\mathrm{pr}(\mathrm{P}|\mathrm{C})$, is high, and prototypical properties in the narrow sense (ins), for which additionally the reverse conditional probability, $\mathrm{pr}(\mathrm{C}|\mathrm{P})$, is high. I term the former frequent and the latter diagnostic properties and I will argue that a property's typicality contribution is best quantified with subjective probability estimations of its diagnosticity for and probability in the category. Frames enable a fine-grained representation of conceptual structure and are therefore a suitable representation format for this purpose. The main hypotheses investigated in this thesis are the following:

1) The typicality of SCs for C is determined by the probability that the SCs have prototypical properties P of C.
2) Prototypical properties are those properties brought about in C's evolutionary history that

a) are possessed by SCs with high statistical frequency, $\mathrm{pr}(\mathrm{P}|\mathrm{C}) = \mathrm{high}$, allowing for the inference $\mathrm{C} \Rightarrow \mathrm{P}$, termed frequent or prototypical in the wide sense (iws),
b) are additionally discriminative with regard to contrast categories of C, $\mathrm{pr}(\mathrm{C}|\mathrm{P}) = \mathrm{high}$, allowing for the inference $\mathrm{P} \Rightarrow \mathrm{C}$, termed diagnostic or prototypical in the narrow sense (ins).

3) Humans build up a conceptual structure which incorporates subjective estimations that are based on the observation of these probabilities, formalised as $\Pr(\mathrm{P}|\mathrm{C})$ and $\Pr(\mathrm{C}|\mathrm{P})$.
4) The structure of prototype concepts is best described by frames. A frame has several attributes $\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_\mathrm{n}$. Each attribute A_i has several possible values $\mathrm{V}_{\mathrm{i}1}, \mathrm{V}_{\mathrm{i}2}, \ldots, \mathrm{V}_{\mathrm{im}_\mathrm{i}}$. Properties correspond to values, property dimensions to attributes. The attributes and values receive weights that reflect their typicality contribution in the following way:
   a) probability: reflects prototypicality iws, estimated for each V_{ij}. The attribute similarity of a SC to C is the sum of the minimum value probabilities for each attribute value:
   $$\mathrm{Sim}(\mathrm{SC}, \mathrm{C}|\mathrm{A}_\mathrm{i}) = \sum_{\mathrm{j}=1}^{\mathrm{m}_\mathrm{i}} \min\left(\Pr\left(\mathrm{V}_{\mathrm{ij}}|\mathrm{C}\right), \Pr\left(\mathrm{V}_{\mathrm{ij}}|\mathrm{SC}\right)\right)$$
   b) diagnosticity: reflects prototypicality ins and is best applied to attributes. Defined as the maximum of the reversed value probabilities per attribute $\Pr\left(\mathrm{C}\middle|\mathrm{V}_{\mathrm{ij}}\right)$ and normalised by the sum of these maxima
   $$\mathrm{diag}_{\mathrm{A}_\mathrm{i}}(\mathrm{C}) = \frac{\max\left(\Pr(\mathrm{C}|\mathrm{V}_{\mathrm{i}1}), \ldots, \Pr(\mathrm{C}|\mathrm{V}_{\mathrm{im}_\mathrm{i}})\right)}{\sum_{\mathrm{i}=1}^{\mathrm{n}} \max\left(\Pr(\mathrm{C}|\mathrm{V}_{\mathrm{i}1}), \ldots, \Pr\left(\mathrm{C}\middle|\mathrm{V}_{\mathrm{im}_\mathrm{i}}\right)\right)}.$$
5) The typicality of SCs can be determined by comparing their value probability distributions with the prototype distribution, weighted by the attributes' diagnosticity, as the diagnosticity-weighted average of similarity according to the formula:
   $$\mathrm{typ}(\mathrm{SC}, \mathrm{C}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{diag}_{\mathrm{A}_\mathrm{i}}(\mathrm{C}) \cdot \mathrm{Sim}(\mathrm{SC}, \mathrm{C}|\mathrm{A}_\mathrm{i})$$

While it is assumed by many that category prototypes are identifiable, specific accounts of the identification of typical properties and parametrical representation of prototypes that predict typicality are rare. Two remarkable exceptions are the family resemblance score (Rosch & Mervis, 1975), which represents prototypes in property lists with application scores that reflect each property's frequency in the category, and a frame-adapted version of Tversky's contrast model (Smith et al., 1988), which represents prototypes in frames that quantify each value with its frequency in the category, measured as number of votes, and diagnosticity weights derived from the frequencies of the values in the category and in the contrast category. I will show that the probabilistic prototype model makes typicality predictions that are correlated with mean rated typicality of the same magnitude as the other models. Furthermore, they have the advantage of only relying on property generation data and probability estimations, without employing the unclear and possibly biased notions of property applicability and number of votes. Conversely, probability ratings can be easily used to represent the variables in the other

models and then predict typicality with correlations in the same order of magnitude as the other models.

Many promising accounts for various phenomena related to prototype theory have been brought forward. This thesis aims at contributing a way to find probabilistic representations of category prototypes that provide a solid foundation for further work. It is a contribution to the extension of frame theory developed in the CRC991 "The structure of representations in language, cognition and science" (e.g., Petersen (2007), Votsis, Schurz (2012), Löbner (2014), Kornmesser and Schurz (2020)) as well as to research questions on the structure and content of mental representations.

Concepts are a topic of cognitive science and work on them is part of an interdisciplinary investigation involving disciplines such as philosophy, linguistics, psychology and information technology. While probabilistic prototype frames are a normative account of concepts and as such a philosophical contribution to the field, the use of psychological and statistical techniques for their empirical verification is indispensable and critically reflected upon throughout this thesis.

Content

1 Introduction: The content and format of concepts

In this chapter, I introduce the basic notions used in this thesis. In section 1.1, a distinction between concepts, categories and words is made. In section 1.2, I compare frames and property lists, the most important formats for representing conceptual content. In section 1.3, the main theories of conceptual content are compared: the classical theory, prototype theory, exemplar theory and theory theory. Section 1.4 presents the outline of this thesis.

1.1 Concepts, categories and words

Concepts are mental representations of objects and events in the world and as such also called the building blocks of thoughts. The represented objects and events can be very concrete and specific representations of individuals, like *Aristotle*, or they can be more abstract and refer to sets of objects and events like *human*. Concepts of the latter kind will be called category concepts and the sets of objects or events they refer to will be called categories. Whenever the distinction is important, concepts will be written in italics (the concept *tree*), the category they refer to unformatted (the category tree) and the symbol that represents the concept/category in language in quotation marks (the word "tree").

The semiotic triangle (Figure 1a), first published in "The Meaning of Meaning" by Ogden and Richards (1923), shows the relationship between words, concepts and categories: concepts (thoughts) refer to categories or concrete entities in the world (referents) and words (symbols) symbolise or stand for concepts. Meaning is the relationship between words and referents (ibid., p. 11). In my thesis, I propose a way to represent concepts that refer to categories from a specific domain: common-sense categories. In this domain, we are usually only interested in a part of the actual concept, namely that part that is present in all competent speakers who know how to use the word correctly. This is illustrated in Figure 1b), a more specific version of the semiotic triangle for common-sense concepts with the specifications I propose in this thesis: they represent the common part of the mental representations of different individuals. These summary representations of individual mental representations are theoretical generalisations whose goal is to make these common parts of meaning accessible. As Cohen and Murphy (1984) note, they are "twice removed from the environment: they are (theoretical) representations, R_1, of (mental) representations, R_2, that in turn represent the external environment D" (p.31). Figure 1b) also illustrates how I view them: they consist of objects, in the example different trees, with common attributes, like LEAF-SHAPE, that have a probability distribution over the different values, like Oval or Circular, depending on how many objects have the specific value. I assume that estimations of these objective probability distributions are stored in combination with the values in the mental representation of the category. In short, probabilistic prototype frames represent common-sense concepts, like *tree*, in terms of properties and their estimated probabilities, and refer to categories of objects in the world, trees, and the objective probability distributions of their properties.

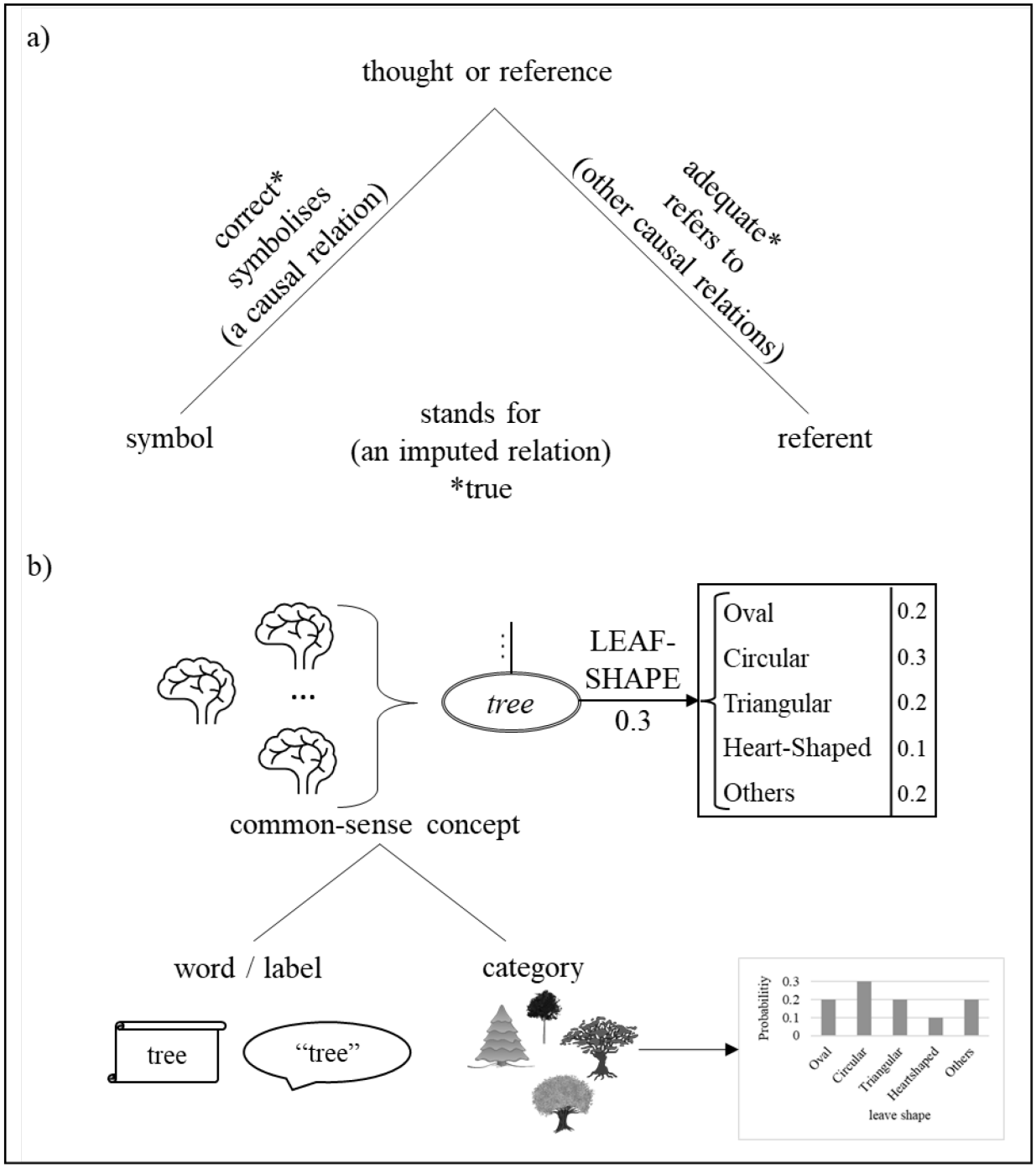

Figure 1: Semiotic triangle a) in the version of (Ogden & Richards, 1923, p. 11) and b) with the specifications proposed in this thesis with an example for trees.

1.2 The representation format of concepts

An analysis of the content of concepts requires a method to represent their components and structure. This section compares feature lists and frames which are the most-common representation formats of conceptual content.

Contrary to the atomistic view of meaning, according to which the meanings of concepts are unanalysable units, decompositional theories assume that the meaning of concepts is determined by properties of the categories they refer to. To make these properties accessible to an analysis, they can be listed in *feature lists*[1], either determined in an analysis, in lexical semantics, or produced by participants in experiments, in cognitive science. An example for a partial feature list for *tree*, including growing conditions and the colour and shape of leaves, is shown in Table 1. An important shortcoming of feature lists is that they do not show how the properties are related: the properties *has green leaves* and *has brown leaves* as well as *has oval leaves* and *has circular leaves* are more closely related to each other than to *grows in forests* and *grows in parks*, which cannot be seen a priori from the list.

Table 1: Example for a partial feature list for tree.

tree
• has green leaves
• has oval leaves
• grows in forests
• has brown leaves
• has round leaves
• grows in parks

For a more fine-grained representation of concepts based on property dimensions, frames are an ideal tool. They overcome this problem by analysing concepts as systems of attributes with admissible values and are a common knowledge representation format in artificial intelligence, linguistics and cognitive science. Minsky (1975) introduced them because other representation formats seem "too minute, local, and unstructured to account […] for the effectiveness of common-sense thought" (p.1). Their usefulness for the analysis of conceptual representations in cognitive science was highlighted in a seminal article by Lawrence Barsalou in which he argues that they "provide the fundamental representation of knowledge in human cognition" (Barsalou, 1992, p. 21). This hypothesis is called the frame hypotheses (FH):

(FH) There is a uniform structure of representations in human cognition, and this structure is essentially Barsalou frames. (Löbner, 2014, p. 64)

Frames organise properties by attributes or property-dimensions which can take several values, the properties. Attributes will in what follows be written in small caps, values capitalised, and the concept to which they apply in lower cases. The thought "This leaf i is green" is thus expressed as $\text{COLOUR}(leaf_i) = \text{Green}$. They are recursive, which means that each value and each attribute are concepts as well and can be further specified, for example $\text{SHADE}(\text{COLOUR}(leaf_i))$ = dark expresses that the leaf's colour has a dark shade. In addition, structural invariants between attributes and constraints between values can be embedded to reflect relationships between attributes and between values, for example the fact that the value

[1] The words feature and property are often used synonymously to refer to qualities or characteristics of things. In the empirical literature, the word feature is predominantly found, while in the philosophical literature the word property is more common.

Oval occurs only together with certain attribute values like Green or Brown on the COLOUR attribute or that having a value on the COLOUR attribute requires having a value on the SHAPE attribute.

Figure 2 shows an example for a partial frame for *fish* that illustrates the components of frames and their recursivity. It represents the properties of fish to have round eyes, to live in water and to move by swimming. These properties can be functionally expressed as

$$\text{VISUAL-SENSE-ORGAN}(\text{fish}) = \text{Eyes},$$
$$\text{SHAPE}(\text{VISUAL-SENSE-ORGAN}(\text{fish}) = \text{Eyes}) = \text{Round},$$
$$\text{LIVING-ENVIRONMENT}(\text{fish}) = \text{Water},$$
$$\text{MODE-OF-LOCOMOTION}(\text{fish}) = \text{Swim}.$$

The *fish* frame contains a structural invariant between the MODE-OF-LOCOMOTION-attribute and the LIVING-ENVIRONMENT-attribute as a secondary predicate:

$$\text{restricts}(\text{living-environment}, \text{mode-of-locomotion}),$$

reflecting the fact that not all modes of locomotion are possible in all living environments (swimming is not possible in the air and flying is not possible in the water). On the value level there is the constraint

$$\text{requires}(\text{MODE-OF-LOCOMOTION}(\text{fish}) = \text{Swim}, \text{LIVING-ENVIRONMENT}(\text{fish}) = \text{Water}),$$

reflecting that water is required for swimming.

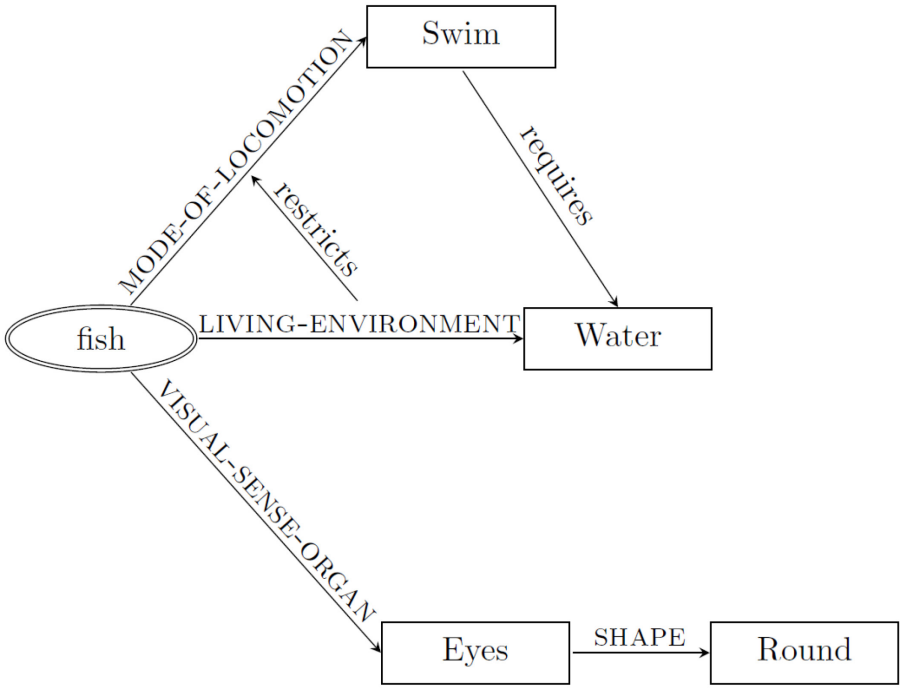

Figure 2: Example for a partial frame for the concept fish.

Frames provide more structure to conceptual representations than feature lists: they structure properties into attributes with different values. Instead of just storing properties like *is brown*, *is green*, *is round*, and *is oval*, these properties are structured in a COLOUR and SHAPE attribute. This way of thinking is appealing from a logical perspective, as pointed out in Smith et al. 1988 (p. 487):

> Without the notion of an attribute, how can one ever know that a blueberry is an instance of *nonred fruit*? To know that blue counts as *nonred* while round does not, one must know that a certain set of values (the colors) constitutes an attribute.

Empirically, attribute-value structures are supported by the observation of an increase in reaction time for attribute shifts in comparison to value shifts (participants switch faster from identifying red objects to identifying blue objects, than from red objects to round objects), for humans and other animals. This indicates that structuring properties along dimensions is a fundamental property of cognitive processing. (Barsalou, 1992, p. 26).

Different versions of frames can be found in the literature. Figure 3 compares a Barsalou frame with the version used in this thesis. In Barsalou frames, attributes are aspects of a concept and values are types of these aspects. For the concept *tree*, LEAF is an aspect of *tree* and COLOUR and SHAPE are aspects of LEAF, which also illustrates the recursivity of frames. *Oval* and *Round* are types or values of the SHAPE attribute. In the frames used in this thesis, attributes are functional and represented as labelled edges. Values or collections of admissible values are represented in nodes. This is a more compact way of presenting the same information and it is simple to add additional information to the values, like probability information.

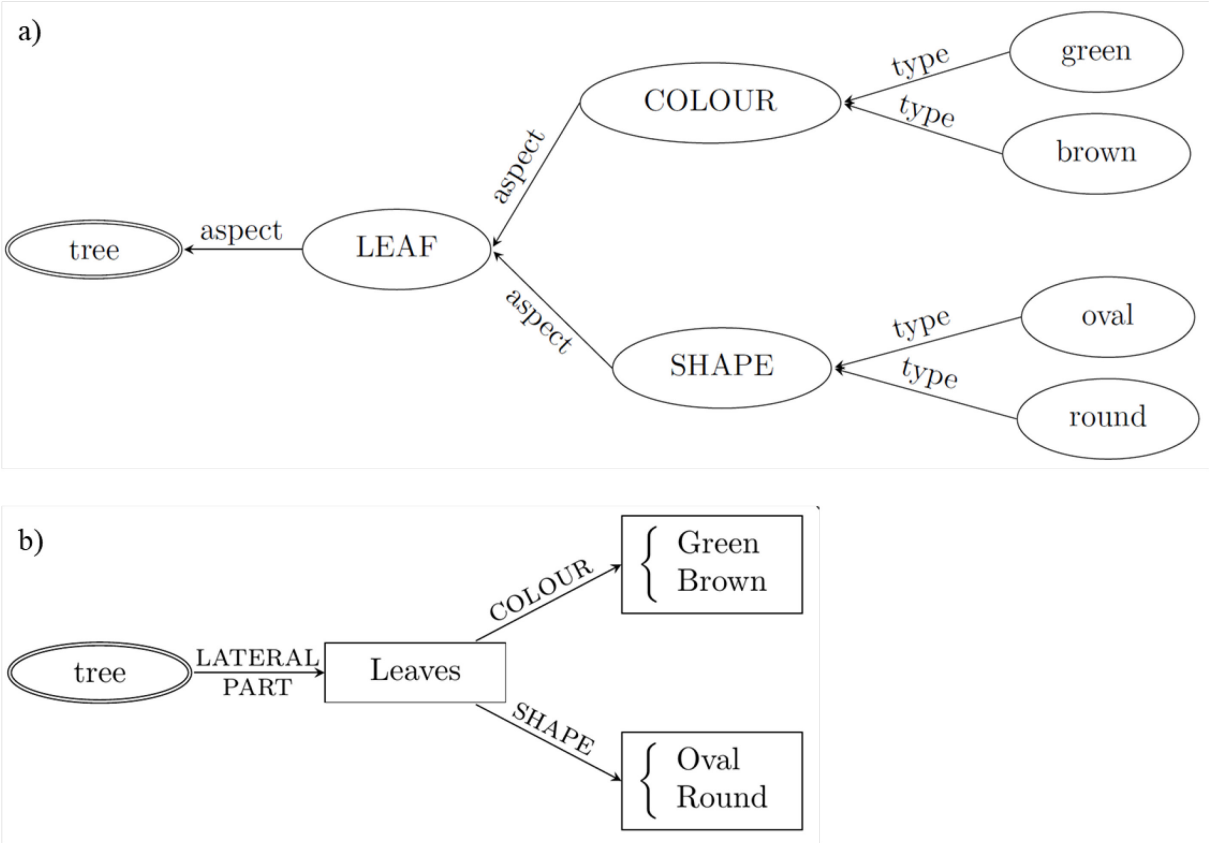

Figure 3: The concept tree represented in a) a Barsalou-frame, b) a frame in the style of this thesis without probabilities.

In the collaborative research centre CRC991 "The Structure of Representations in Language, Cognition, and Science", located at the Heinrich-Heine-University Düsseldorf from 2011 to 2020, Barsalou frames were extended in several directions. The main change that applies to most of their subtypes, summarised as Düsseldorf frames, is to include only functional attributes and to be typed with types which are hierarchically related. These types can add additional constraints on the values an attribute can take. Frames were shown to be an adequate tool to model various linguistic and philosophical aspects of meaning, like different word types

(Petersen, 2007), sentences (e.g. Kallmeyer & Osswald, 2013), dynamic events (Balogh & Osswald, 2021), discourse (Berio et al., 2017), uncertainty (Sutton & Filip, 2017), scientific theory change (Schurz & Votsis, 2014) and default inheritance (Strößner & Schurz, 2020). Their fine-grained structure offers a simple way to point out which node is exactly targeted by a modifier (e.g. Goldschmidt et al., 2017). Dynamic events like the drying of a shirt or the rising of a balloon have also been represented in frames with an update operation (e.g. Osswald & van Valin, 2014). Different aspects of granularity are shown in a combination of frames called cascades (Löbner, 2021). In Schuster et al. (2020), we present ways to embed stochastic information in frames. Frames become a particularly powerful representation format when they are enriched with stochastic information and this kind of frame is presented in detail in section 4.3.

An important distinction is to be made between instantiated and uninstantiated frames, which are used to represent individual and sortal concepts respectively in the terminology of Petersen (2007). Fully instantiated frames represent single events or entities. Uninstantiated frames represent categories of events or entities. In them, not all attribute values are specified. This is illustrated in Figure 4 for the values of the COLOUR attribute in the completely uninstantiated frame for *fruit,* which contains information on all possible values for fruit, then for *apple* as subcategory, which is partially instantiated and has a narrower range of admissible values, and the frame for a specific $apple_i$, which has one specified value, a fully instantiated COLOUR-attribute.

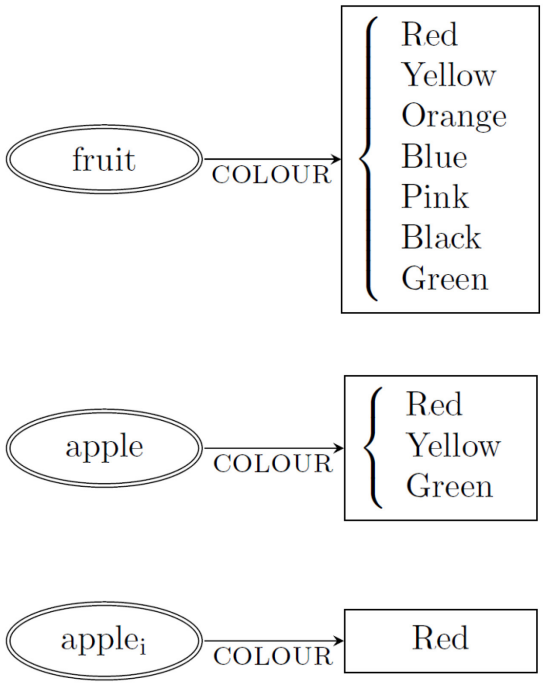

Figure 4: Three levels of abstraction/generality for the COLOUR-attribute for the categories fruit and apple and one specific apple i.

A special kind of uninstantiated frames are classificatory frames, as used in Chen (2003), Schurz and Votsis (2014), Votsis and Schurz (2012) and Kornmesser and Schurz (2020), in which subcategories are defined by assigning them values from the frames for categories. An example for this kind of frame is presented and explained in section 2.1 (Figure 6).

1.3 The content of concepts

This section presents the main decompositional theories of meaning: the classical theory, the prototype theory, the exemplar theory and the theory theory. Their main stances are summarised in Table 2. The view of the classical theory, according to which meaning is a list of singly necessary and jointly sufficient conditions, was challenged by the other theories. For prototype and exemplar theories, properties that contribute to meaning do not have to be necessary and jointly sufficient. Instead, they focus on characteristic properties either of the category as a whole (prototype theory) or of category members (exemplar theory). The theory theory takes this view to be too restricted and emphasises the role of background knowledge and inferential processes, which makes meaning "an explanatory principle common to all category members" (Murphy & Medin, 1985, p. 289). The conditions from classical theory define category membership and thereby its extension clearly: any entity that fulfils the conditions is an equally representative member of the category. Empirical findings indicate that not all category members are perceived as equally representative. This led prototype theory and exemplar theory to base categorisation on similarity, either to the category prototype or to other category members. This makes the extension of categories fuzzy, as there are often borderline cases which are perceived as members by some and as non-members by others. This problem is solved in theory theory in which not only similarity, but also inferential processes constitute meaning. They can dissolve borderline cases and make the categories' extension clear. It is however not clear whether fuzziness is a problem that has to be solved or simply a phenomenon that occurs with certain concepts. While classical and theory theory are not able to describe this phenomenon, prototype and exemplar theory offer an account of borderline cases. In classical theory, conceptual composition is accounted for by combining the necessary and jointly sufficient conditions of the words that are combined. In theory theory, it is explained with the use of inferential processes and background knowledge considerations. Prototype theory and exemplar theory do not have inherent mechanisms to explain conceptual combination. There are, however, developed accounts for conceptual combination with prototypes by James Hampton (1987, 2007) and Smith et al. (1988). While both classical theory and theory theory have straightforward ways to determine the meaning in conceptual combination and the extension of categories, they lack plausibility for natural language: there is empirical evidence that people do not perceive categories to have clear extensions and also that conceptual combination needs to take into account the graded category structure that is the basis of prototype and exemplar theories. The fact that they incorporate this empirical evidence makes them more plausible accounts of natural language.

Table 2: Theories of concepts' stances on meaning, categorisation, extension and conceptual combination.

theory	meaning	categorization	extension	conceptual combination	plausibility for natural language
classical	singly necessary and jointly sufficient conditions	all or none	clear	explained	not plausible
prototype	characteristic properties	similarity	borderline cases	with additional assumptions	plausible
exemplar	properties of category members	similarity	borderline cases	not explained	plausible
theory	explanatory principle	similarity + inferential processes	clear	explained	not plausible

Each theory of concepts has its advantages and disadvantages, which led to the development of different forms of conceptual pluralism. Weiskopf (2009) argues that the context determines which theory of concepts is best suited to account for the phenomenon in question and Schurz (2012) argues that it is the concept's domain. I take both assumptions to be true and think that the content of a conceptual representation depends on the goals set by the context which is often determined by the concepts' domain. If the goal is for example legal adequacy, a classical definition should be the content of choice because it offers a straightforward way to determine extensions. In the domain of legal concepts, they are furthermore readily available most of the time. If the goal is scientific accuracy, the background theory needs to be taken into account. This applies for the domain of scientific concepts. If the goal is cognitive plausibility and the domain is common-sense concepts, definitions are underinformative because they do not incorporate characteristic properties, and inferential processes as well as explanatory principles required by the theory theory are difficult to identify. They are best accounted for by the similarity-based theories. The aim of this thesis is to find a cognitively plausible representation of the meaning of common-sense concepts. For this goal and in this domain, prototype theory is the best choice. The prototype frames that are proposed in what follows can however incorporate properties of the classical theory and theory theory in the following way: properties that are necessary for category membership can receive a weight that is greater than the sum of any non-necessary properties, so that it cannot be compensated for by them. Or they are connected to other properties in a way that the negation of these properties leads to the loss of many others. Such and other relations between properties and property dimensions can be added, and when necessary quantified, as constraints, to account for the relations between properties like in theory theory.

1.4 Thesis outline

Having laid the foundation in the last sections, I will proceed as follows. In chapter 2, I present the history of the prototype theory of concepts, discuss its relationship to classical definitions and present experimental results that show its non-classical properties. Chapter 3 is an in-depth discussion of typicality as the theory's empirical basis with a special focus on typicality orderings, whose explanation, prediction and implications for concepts are the goal of this thesis. I compare different studies in a meta-analysis to gain insights into the experimental methods they involve and their intersubjective stability. When light is shed on the research subject, I present a formalisation of prototype theory and discuss the decisions to be made for any quantified prototype model in chapter 4: criteria for the identification of typicality-contributing properties, and for the selection of a representation format and a similarity relation. Chapter 5 then proceeds to develop the probabilistic prototype frame model, based on Gerhard Schurz' work on evolutionary normality. The next chapters describe how I collected and used empirical data to test the developed hypotheses. In chapter 6, I describe the results of three experiments that I designed to collect the subjective probability estimations on which our model is based. How I used this data to predict typicality with probabilistic prototype frames in several versions is presented in chapter 7. In chapter 8, I investigate how well the two other accounts, the family resemblance score and the contrast model, fare when applied to new data and when probability data is used as their input. Chapter 9 is the conclusion in which I summarise the results and give an outlook on promising future research.

2 The prototype theory of concepts

This chapter first provides an overview of the historical development of prototype theory and its main theses, empirical findings and criticisms (section 2.1). Then, it contrasts prototype theory with the classical theory of concepts (section 2.2). Section 2.3 summarises the logical properties of prototypes, focusing on category membership (2.3.1), logical operators (2.3.2), default inheritance (2.3.3) and conceptual combination (2.3.4). Section 2.4 summarizes the foregoing sections.

2.1 Historical overview

The research on prototype theory began with the finding presented in Berlin and Kay (1969) that colour space is not divided arbitrarily in different languages and cultures, but is universally structured around certain clear examples. Eleanor Rosch, in the beginning under the name Rosch Heider, built up on these findings and confirmed and extended them for young children (Rosch Heider, 1971), New Guinea Dani, a people with undifferentiated colour terminology (Rosch Heider, 1972), and for shapes (Rosch, 1973a, 1975a). She generalised these results to semantic categories like furniture or vegetables (Rosch 1973b, 1975b) and summarised and analysed her findings in Rosch (1978), where she states that human categorisation is guided by two basic principles: cognitive economy and the perception of a correlational structure in the environment. She describes two important dimensions of categories: vertical and horizontal. Figure 5 illustrates these dimensions with some examples. The vertical dimension generates a taxonomy of objects where the inclusiveness of the category is greater the higher a category is in the taxonomy – *living thing* for example includes many more objects than *collie.* Accordingly, the number of properties shared between category members is lower, the higher the category is in the taxonomy: *living things* share only few unspecific properties like *being alive* and *being material*, while *collies* share many specific properties like *medium-sized* and *pointed snout.* Typicality effects are found on the horizontal dimension. There is one special category level in which categories include many objects and have many common properties. She calls categories on this level basic level categories, and states that they are on "the most inclusive (abstract) level at which the categories can mirror the structure of attributes perceived in the world" (p.5). The supercategories of basic-level categories, like *mammal*, *vehicle* and *furniture*, also have a special role, because they "are sufficiently abstract that they have few, if any attributes common to all members" and "consist almost entirely of items related to each other by means of family resemblances of overlapping attributes" (Rosch & Mervis, 1975, p. 576). It is on this level that typicality orderings are the most researched and pronounced.

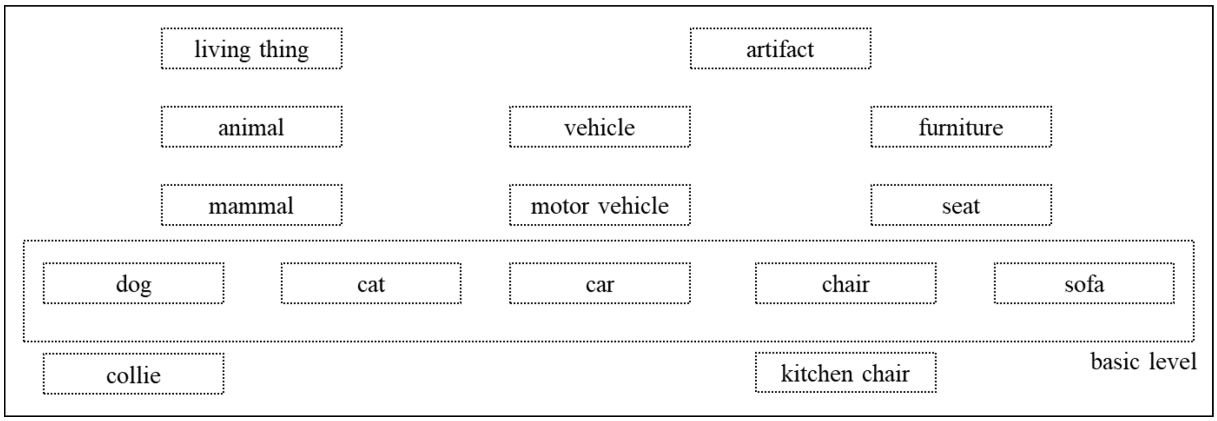

Figure 5: Vertical and horizontal category structure described in Rosch (1978) with her examples (p. 5).

Figure 6 shows how Kornmesser and Schurz (2020) represent the core properties of the vertical dimension of prototype theory in a theory or classificatory frame and how some of the main theoretical concepts from prototype theory, *superordinate category, basic level category,* and *subordinate category*, derive from this core. Superordinate categories are the supercategories of basic-level categories, like *mammal* for *dog*. They include many SCs and are associated with few common properties, few common motor movements and low shape similarity. Basic-level categories also include many SCs but are at the same time associated with many common properties, many common motor movements and a high similarity of shape. Subordinate categories are the SCs of basic-level categories, like *collie* for *dog*. They include only few SCs, which share many properties and motor movements and have a high similarity of shape. The basic level thus maximises the level of inclusion as well as the number of common properties and motor movements and the extent of shape similarity.

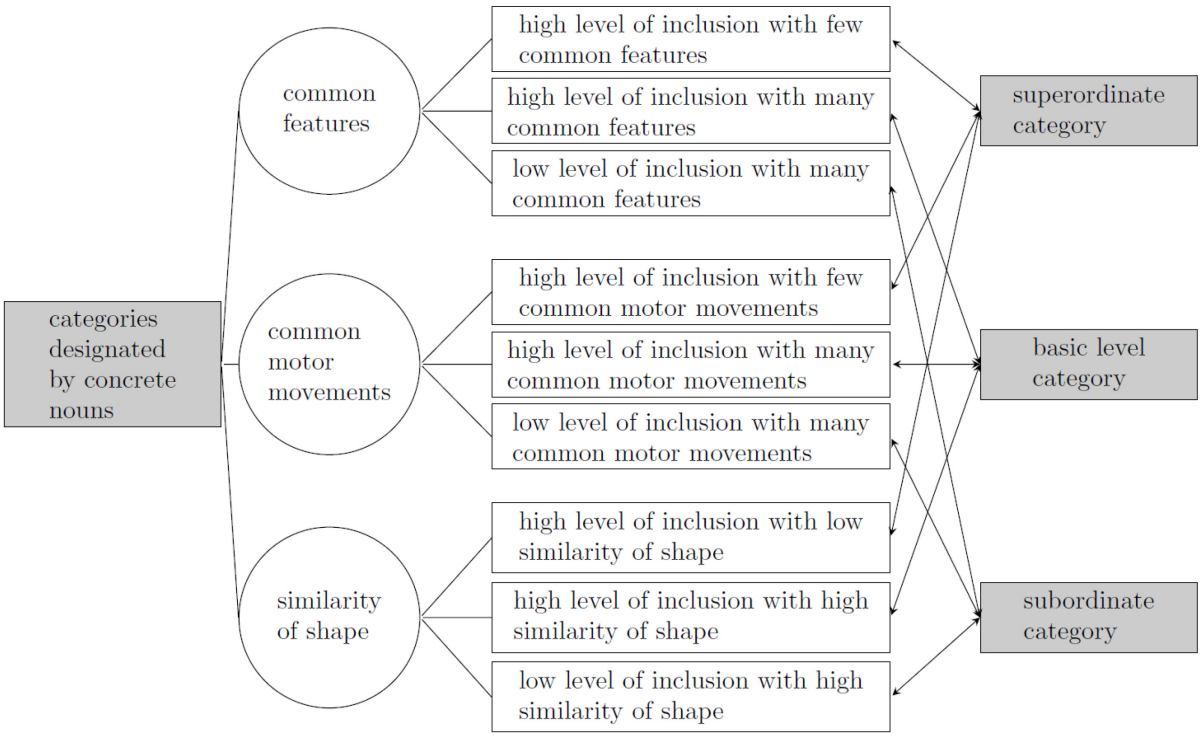

Figure 6: Theory frame representing the core of the vertical dimension of the prototype theory. Adapted from figure 6 in Kornmesser & Schurz, 2020, p. 1324

The early prototype theory treated prototypes as real category members, namely those with the highest typicality ratings: "[b]y prototypes of categories we have generally meant the clearest cases of category membership defined operationally by people's judgments of goodness of membership in the category" (Rosch 1978, p. 12). She points out that

> "To speak of *a prototype* at all is simply a convenient grammatical fiction; what is really referred to are judgments of degree of prototypicality.[…] For natural-language categories, to speak of a single entity that is the prototype is either a gross misunderstanding of the empirical data or a covert theory of mental representation." (Rosch, 1978, p. 15)

The interpretation of prototypes as summary representations of properties instead of real category members began with the article "Family Resemblances" by Rosch and Mervis (1975). They present an analysis of the property structure of categories, based on Wittgenstein (1953)'s observation that for many categories, like *game*, category members lack defining properties common to all of them. Instead, category members share different subsets of properties, like the members of a family who do not look identical but resemble each other. Some but not all games are played with a ball or played for fun. Rosch and Mervis found that typical category members share a lot more properties than atypical ones and that the typicality ordering is highly correlated with the amount and centrality of the properties they share, which they calculated as a family resemblance score (see section 8.1). James Hampton (1979) found that the distinction between defining properties (present in all category members) and characteristic properties (present in some of the category members) makes no difference for the prediction of category membership. He introduces the notion of a *polymorphous concept* (going back to Gilbert Ryle, e.g. Urmson, 1970), a concept "in which an instance is classified […], if and only if it possesses at least a certain number of a set of features, none of which need to be necessary or sufficient in itself." (p. 450) and notes the close relation to Rosch and colleagues' work, which he develops by giving a methodology to identify prototypical properties and by showing that the extent of SCs having these properties is correlated with category membership decisions. In the following years, Hampton (1982, 1988a, 1988b, 1997b) collected evidence that common-sense categories such as fruit or furniture are handled differently than classical logic would predict (see section 2.3), which indicates the classical theory is not a suitable tool for describing human reasoning with categories in this domain.

McCloskey and Glucksberg (1978) present evidence that SCs with medium typicality show a high amount of inter- and intra-subjective variability and conclude that "these data suggest that natural categories are fuzzy sets, with no clear boundaries separating category members from nonmembers" (p. 462). Hampton (1979) confirmed these results. The vagueness of category borders is another important phenomenon of prototype theory, discussed in section 2.3.1.

In Barsalou (1983) and (1985), further evidence for the intersubjective stability and cognitive reality of typicality orderings was presented. He presents typicality orderings for "ad hoc categories" and "goal-derived categories", which are not based on property clusters in the environment and also probably not integral parts of memory, like *things-to-take-from-one's-home-during-a-fire* and *things-to-eat-on-a-diet*, and concludes that "this appears to be the result of a similarity comparison process that imposes graded structure on any category regardless of type" (Barsalou, 1983, p. 211).

The most common objections to prototype theory are that not all categories have a prototype and that prototypes are not compositional. Fodor (1981) argues that complex phrases do not have prototypes: "There may, for example, be prototypical *cities* (London, Athens, Rome, New

York) […] but there are surely no prototypical *American cities situated on the East Coast just a little south of Tennessee.*" (pp. 296-297). Osherson and Smith (1981) and Fodor and Lepore (1996) point out that the meaning of noun-noun-composita cannot be explained by prototype theory. This objection is called the pet-fish-problem, because a famous example is that the prototype of *pet fish* is neither a prototypical *fish* nor a prototypical *pet*. It is discussed in section 2.3.4.

These authors question the usefulness of prototype theory if it can neither explain the content of all concepts, as there are many which do not seem to have a prototype, nor explain the compositionality of concepts. Schurz (2012) argues that prototype theory can only be applied in the domain that it aimed to describe in the first place: categories that are based on the correlational structure of the environment. He sees their role not in giving fool-proof definitions, but in enabling "fast and efficient predictive and diagnostic reasoning" (p. 546). It is thus not necessary to assume that prototypes explain concepts from all domains and all related phenomena for them to be an integral part of our cognitive system and as such a worthy research topic. His account of prototype theory is the basis for the prototype frames proposed in this thesis and developed in chapter 5.

Another popular argument against giving typicality an important position in meaning contribution is that it is also found in categories that are clearly well defined, like *even-number.* Armstrong et al. (1983) conclude from their demonstration of typicality orderings in well-defined categories that prototypical properties cannot be the core of mental representations and propose a dual description of concepts, in which classical definitions are the core that determines category membership and prototypical properties are the periphery which aids the identification of category members. Larochelle et al. (2000) show that the typicality orderings in well-defined categories are different from those in the prototype domain: the ordering disappears if productive frequency and familiarity are controlled, which is not the case for prototype categories which will be discussed in section 3.3.1. Schurz (2012) points out that prototype reasoning, due to its unconscious nature, is sometimes applied to well-defined categories. But in these cases, it is no longer efficient and can be faulty.

Despite the objections, prototype theory stayed an important topic in cognitive science, linguistics and artificial intelligence. It is widely accepted that many words from everyday language are better described and analysed based on characteristic category properties.

In the beginning, it was assumed that prototypes have average values on the dimensions important for the category, which is shown for example in the quote "centrality shares the mathematical notions inherent in measures like the mean and mode. Prototypical category members have been found to represent the means of attributes that have a metric, such as size" (Rosch, 1978, p. 12). This view is too restricted. Barsalou (1985) found that some categories are ordered according to their distance from an extreme value, like *food to eat on a diet,* where the most typical examples are those most close to the extreme of having 0 calories, and not those with a medium amount of calories. Rosch pointed out already that "both representativeness within a category and distinctiveness from contrast categories are correlated with prototypicality in real categories" (1978, p.13). This was confirmed by Ameel and Storms (2006), who found that in geometrical models of categories, typicality was better predicted if the reference point was not the geometrical centroid but instead moved as far as possible from the contrast category's members. Further evidence for the importance of the contrastiveness of prototypes is presented in Douven (2019) for colour prototypes. Rosch herself points out in

later work that understanding prototypes as averages is too narrow and instead, depending on domains, different factors determine prototypes. She criticises that prototypes are often equated with averages: "Unfortunately, the psychological research community has largely come to take the word prototype as having this and only this meaning" (Rosch, 2011, p. 101). She lists the following possible determiners of typicality orderings (pp.101-103):

- the most salient ones for sensory domains like colour and shape,
- the statistical average / family resemblance ones,
- ones that reflect ideals as the extreme instead of the average[2],
- stereotypes of social roles, where idealisations determine the prototype (president, mother, bus driver, …),
- stereotypes of groups of people (race, ethnicity, gender, …)
- precedents, for example in law where judgements are made based on preceding judgments in similar cases,
- prototypes determined by goals (*foods to eat on a* diet) or for ad hoc categories (*things to take from a burning house*) (Barsalou (1983)),
- reference points in formal structures (*multiples of 10*),
- "easy-to-understand or easy-to-imagine examples illustrating abstract principles" (p. 102), like examples used to explain scientific concepts,
- prototype as a good fit with causal theories,
- prototypes determined by the earliest or most recent experience with an item, like doctors' tendentially using their most recent cases for diagnosis (Brooks et al., 1991),
- prototypes as a list of some of the best examples, like apple-orange-banana meaning *fruit* in American sign language (Rosch et al., 1976),
- prototypes determined by strong personal experience with category members,
- "Idiosyncratic prototypes generated at a particular moment as a result of the confluence of an individual's past experience and present circumstances" (Rosch, 2011, p. 103).

What unites these different kinds of prototypes is that they all serve as cognitive reference points, which is also the title of one of Rosch's first papers on the phenomenon. The typicality ordering results from the similarity of different subcategories to the category's prototype, and it does not matter which of these principles led to its construction.

[2] "For American undergraduates, the best examples of cities ("Now that's a real city!") are the largest and most cosmopolitan, such as New York, Paris, and London, not average cities." (Rosch 2011, p. 101)

2.2 Relation to classical definitions

Prototype theory was developed to account for the observation that category concepts are associated with typicality orderings. As definitions cannot distinguish between category members, this suggests that certain properties are important parts of category concepts, despite being neither necessary nor sufficient for category membership. Figure 7 compares how the words *bachelor, fruit* and *molecule* can be defined in frames with properties fulfilling the criteria of classical theory (using definitions from the WordNet[3]) and examples for ones from prototype theory. In classical theory, a bachelor is a man (a person with the value Male on the GENDER attribute and the value Adult on the AGE attribute) who is not married (Unmarried on the MARITAL STATUS attribute). The prototype definition reflects that bachelors are usually under 40 and not in a serious relationship. The classical definition of fruit is "the ripened reproductive body of a seed plant". The prototype definition reflects that fruit often taste sweet and sometimes sour, are often consumed in a juice or as dessert and they grow often on trees and sometimes on bushes. The classical definition of molecule is "(physics and chemistry) the simplest structural unit of an element or compound". The prototype definition reflects the common knowledge that molecules are small, often consist of a small number of atoms (as the most commonly known molecules presumably are H_2O and O_2) and that they are material. This example shows that prototype representations are not necessarily scientifically correct but reflect the common ground shared between differently educated individuals.

As already mentioned in section 1.3, prototypes can be seen either as an alternative for classical definitions or as a complementation. Many common-sense categories have no commonly accepted definitions but even if they had, it would be impossible to explain all phenomena observed in the context of concept use with them. A cognitively plausible theory of concepts should include properties that are associated with categories.

[3] www.wordnetweb.princeton.edu, accessed on 29.04.2019.

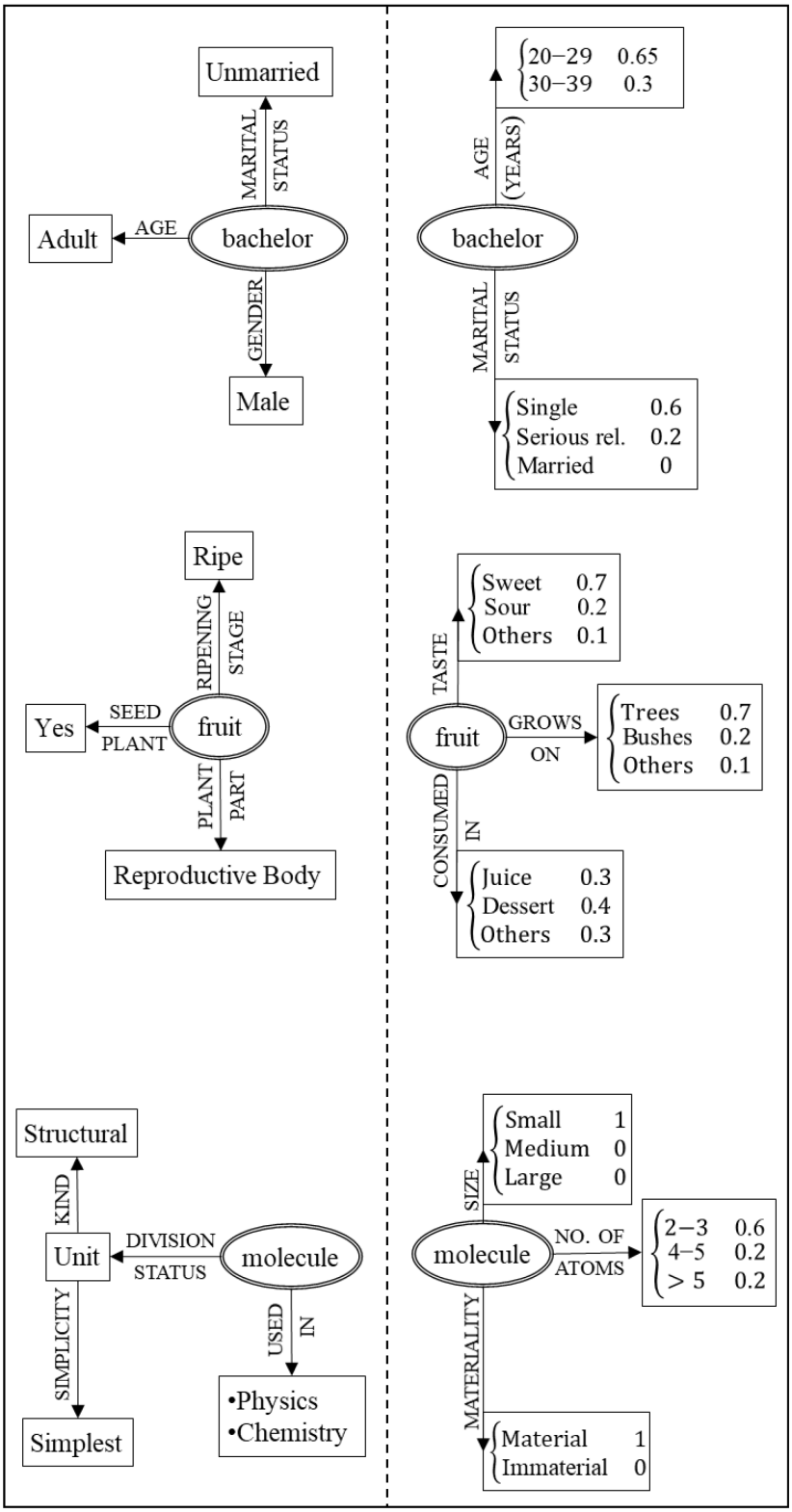

Figure 7: Frames for bachelor, fruit and molecule in classical theory with definitions from www.wordnetweb.princeton.edu (left) and prototype theory with fictional attributes and values (right).

2.3 The (il)logical properties of prototype concepts

This section summarises research on the logical properties of prototype categories. These findings support the psychological reality of prototype representations contrary to classical definitions and they show that finding reliable prototype representations as proposed in this thesis is a valuable endeavour with many practical applications.

As pointed out in section 1.3, the prototype theory of concepts includes no clear criterion for category membership (section 2.3.1). This has implications to how prototype concepts behave when combined with logical operators like conjunction, disjunction and negation (section 2.3.2) as well as to their properties in reasoning patterns (section 2.3.3) and to how conceptual combination can be explained (section 2.3.4).

2.3.1 Category membership

In classical definitions category membership is clearly defined to apply to all entities that fulfil all the singly necessary and jointly sufficient conditions. For prototype theory it is less clear what constitutes category membership. What seems to be clear is that typicality degrees do not correspond to membership degrees, because on the one hand there are atypical members that are clearly members, like *golden chair* for *chair* (cf. Schurz, 2012/2012, p. 541), and on the other hand clear non-members are rated to have some typicality, like *dolphin* for *fish*.

McCloskey and Glucksberg (1978) found that intersubjective agreement about category membership was lowest for medium typicality levels. Hampton (1979)'s analysis of category membership confirmed that there were cases in which membership was not clear, when the percentage of participants that rated the SCs in question to be category members was around 50.

Hampton (2007) points out that questions of category membership are due to the vagueness of concept application, which is the "inevitable result of a knowledge system that stores the centres rather than the boundaries of conceptual categories" (p. 380). In his threshold model, both typicality and category membership result from similarity, because, if they were truly distinct phenomena,

> "typicality T may turn out to be a dimension of purely psychological interest, responsible for the range of typicality effects but of little value for explaining the role of concepts in determining the truth of sentences, while degree of membership M is more a matter of concern for logicians and ontologists." (p. 358).

The threshold model thus maintains that prototypical properties, from which typicality derives, are valuable parts of conceptual representations in general. While typicality is conceived as a linear function that rises with similarity, membership is a smooth function that reaches 1 before typicality is maximal (reflecting that *penguins* and *robins* are both clearly birds, albeit with different typicality) and might be graded in a similarity interval surrounding the membership threshold. He proposes to operationalise it as categorisation probability, which is the percentage of participants who categorise the SC in question as a member. A similar line of thought is found in Shepard's universal law of generalisation (Shepard, 1987), in which the probability of the generalisation of a stimulus depends on its similarity to stimuli observed before.

2.3.2 Logical operators

Hampton (1982) investigated whether common-sense categories (here generalised as A, B, C) follow the principle of class inclusion or transitivity: if A is a member of B and B is a member of C, is A considered to be a member of C as well? In classical definition this is necessarily the case. He found that while most cases followed the principle of class inclusion, there was a non-

negligible percentage of subcategories for which one of the statements was considered false and the other true – A is B, B is C, but A is not C, or A is B, B is C, but A is not C. For example, a bird's nest was rated to belong to the category bed, bed to belong to the category furniture but a bird's nest was not rated to belong to furniture.

In Hampton (1988a), overextension as well as underextension were found when participants were asked about objects being members of A, B and A-or-B, like *hobbies, games* and *hobbies-or-games*. Logically, something that is in A or in B must also be A-or-B and something that is neither in A nor in B cannot be A-or-B. He found examples of SCs being considered a member of A and/or of B, but not of A-or-B (underextension), as well as examples for SCs being neither A, nor B but A-or-B (overextension). *Eating ice-cream* was for example considered a hobby but not a game and also not a hobby-or-game by 50% of the participants. This effect could be explained as a function of the mean membership ratings for A and B of the SCs. He explains the results as effects of joining the properties of both categories: "[w]hen two categories are placed in disjunction, there is a tendency for a mutual interaction of attributes" (p. 588).

In Hampton (1988b), similar results are reported for conjunctions: objects that were rated to be a member of either A or of B but not both, were still counted as members of A-and-B. Logically, an object can only be in A-and-B, if it is a member of both A and B. For example, many participants rated a photocopier to belong to the category office furniture, but not to the category furniture.

Hampton (1997b) researched categorisation behaviour in negated conjunctions, like A, not-B and A-and-not-B. He found that for example *elephant* was not considered a member of *pet* (categorisation probability = .1), clearly a member of *not-birds* (categorisation probability = 1) and also member of the category *pets-that-are-not-birds* (categorisation probability .7).

All these experiments show that the laws of classical logic do not necessarily apply to prototype categories. This strengthens the hypothesis that the meaning of common-sense concepts cannot be cognitively plausibly explained with classical definitions. The fact that there are counterexamples in the prototype domain does however not negate the usefulness of classical reasoning patterns, which were still found to hold in the majority of cases. It does however show that typicality plays a role in reasoning as well and that its effects can overwrite these rules.

2.3.3 **Default inheritance and non-monotonic reasoning**

Default inheritance is an inference strategy in which, in the absence of information to the contrary, properties of C are assumed to hold also for SCs of C. For example, if someone tells me that they saw a bird, I will assume that it could fly, even though I know that there are exceptional birds that do not fly.

In the article "Why stereotypes don't even make good defaults", Connolly et al. (2007) present evidence that the rule of default inheritance is not followed by humans. They found that the rated probability of statements about SCs' prototypical properties was reduced when the SCs are modified with typical (condition B), non-typical (condition C) and two non-typical (condition D) adjectives. An example for the conditions that they presented to their participants is (p.10):

A. Ducks have webbed feet.
B. Quacking ducks have webbed feet.

C. Baby ducks have webbed feet.
D. Baby Peruvian ducks have webbed feet.

The report a mean rated probability on a scale from 0 (not plausible) to 10 (highly plausible) for the A condition of 8.36, for condition B of 7.71, for C of 6.91 and for D of 6.48. This effect of (illogically) reduced likelihood for sentences with modified nouns is called the modifier effect. They conclude that "[t]his set of results is inconsistent with the claim of DS [default-to-stereotype, i.e., default inheritance] that conceptual combination entails the inheritance of stereotypical default values for features that do not figure explicitly in the combination." (p. 13).

There are however different ways to explain the modifier effect. In line with Jönsson and Hampton (2012), in Strößner and Schurz (2020) and Strößner et al. (2021) evidence is presented that the influence of pragmatic effects in default reasoning can be reduced when modified and unmodified sentences are presented in a direct comparison. Then, all possibilities are rated equally likely. Furthermore, they show that the rated likelihood is drastically lower when a modifier is known to be relevant for the ascribed property. Also, in non-monotonic reasoning the inference from a class like "ducks swim" to a subclass like "baby ducks swim" is only valid if it is also known that the class baby is not an exception, i.e., additionally it has to be known "it is not the case that babies do not swim". The reduced likelihood in the example could be explained from the fact that the participants were not sure if baby ducks already have webbed feet or if they develop later (cf. section 5.2).

In Strößner et al. (2021), we analysed the original data that we were kindly provided with and found that in many cases there was a constraint between the modifier and the ascribed property which was hindering the default inheritance in the modified sentences and that in cases where no constraint was present the differences were less grave than reported in the original model. For example, *Bitter nectarines are juicy* might be rated less likely than *nectarines are juicy* because bitter nectarines are usually not ripe and unripe fruit are usually not juicy. Furthermore, in accordance with Jönsson and Hampton (2012), we found that there were only insignificant differences between the C- and D-condition.

While default inheritance is not resulting in the same likelihood for SCs, it is still clearly applied and the reduction in likelihood can be explained by pragmatic effects and constraints imposed by modifiers that are known to be relevant.

2.3.4 **Conceptual combination**

As outlined in the overview, one main problem of prototype theory is its failure to provide a straight-forward mechanism for conceptual combination: the pet-fish-problem or guppy effect. Two prominent solutions were proposed in the literature.

Smith et al. (1988) propose the Selective Modification Model for adjective-noun-compositions which is based on properties represented as attribute-value-structures in a frame, quantified with diagnosticity and frequency weights. In their model, the typicality of a SC for *green vegetable* is for example determined by calculating the similarity of that concept to *vegetable*, then raising the diagnosticity of the COLOUR-attribute and shifting all frequency weights to the Green value. They report high correlations between .85 and .94, except for *white fruit* and *long fruit*, for which most of the chosen subcategories were not typical, as most were neither long nor white. Their method of typicality prediction is discussed in detail and applied in section 8.2.

Hampton (1987, 2007) presents results indicating that prototype combinations are determined by prototypical properties and not by typical category members – to combine *pet* and *fish*, the properties of *pet* and *fish* are combined and not their category members. In his Composite Prototype Model, the properties of both constituents are first pooled and then analysed for interactions, where mutually exclusive properties, like *animate* and *inanimate* for *stone lion*, are analysed in terms of their necessity or impossibility and properties are discarded if it is required to obtain a coherent concept. This new list of properties can prioritise different properties than each constituent category separately and determines the typicality ordering of the conjunction. Similar lines of reasoning can be applied to combine concepts in disjunctions and negations.

In Hampton and Jönsson (2012), a principle for prototype compositionality is presented in accordance with the Composite Prototype Model: "The content of a complex concept is completely determined by the contents of its parts and their mode of combination, *together with general knowledge*." (p. 386). Peter Sutton (2017 and in an unpublished manuscript) presents a way to combine prototype concepts making use of conditional probabilities in Bayesian networks. Strößner (2020a) presents another normative model to account for conceptual combination with conditional probabilities and constraints in prototype frames (see section 9.2.1) building up on this principle.

While prototype theory itself does not include an inherent principle for conceptual combination, it is possible to identify combinatory principles, in particular when prototypes are represented stochastically. The probabilistic prototype frames developed in this thesis are a good basis for a model of conceptual combination with prototype concepts.

2.4 Summary

Prototype theory is an important theory of conceptual representations. Based on perceived differences in the representativeness of SCs for Cs and removing the necessity criterion for properties from definitions, it provides an informative representation that is more in accordance with how cognition works than classical definitions. The observation of borderline cases in category membership, the phenomena of over- and underextension observed in conjunctive and disjunctive reasoning with prototype concepts as well as the employment of default inheritance all further solidify the need for a non-classical way to represent concepts from this domain.

To tackle the more difficult challenges of prototype theory, two strategies have to be employed. First, the domain in which typicality effects are important needs to be restricted to situations in which they serve a purpose, like fast categorisation or reliable common-sense reasoning. Secondly, to account for more complex phenomena, like conceptual combination and reasoning patterns, and to in-depth characterise them, prototypes need to be quantified in order to clearly represent to which degree each of their properties contribute to typicality. The need for in-depth quantified prototype representations guides this thesis.

3 Empirical status of prototype research

In this chapter, I introduce the basic concepts from statistical psychological research that are important for prototype theory (section 3.1), analyse data from existing typicality studies (section 3.2) and related variables (section 3.3). In section 3.4, I summarise and discuss the empirical situation of prototype theory.

3.1 Statistical methods in typicality research

In this section, statistical and numerical concepts and methods used in typicality research are reviewed and critically discussed. I will begin with discussing the role of rating scales in the levels of measurement (3.1.1). Then, I will discuss the measures of central tendency for each level of measurement (3.1.2). The last two sections briefly introduce correlations (3.1.3) and reliability measures (3.1.4).

3.1.1 Levels of measurement

Typicality data is collected in rating scales, a question format which asks participants to rate different items on a scale. The scales are either fully labelled (*Likert* scales) or only at the extremes. Scales can have verbal labels, numerical labels or both. Which kind of data result from rating scales is an ongoing debate. The kind of data determines which statistical procedures are appropriate for their evaluation.

In the 1930s and 40s there was a committee from the British Association for the Advancement of Science which debated the possibility of quantifying human sensations. In an influential article, Stevens (1946) argues that the problem lies in the semantics of the term "measurement". He proposes to follow N. R. Campbell from the committee and define measurement broadly as "the assignment of numerals to objects or events according to rules" (p. 677). This clearly applies to rating scales. More fine-grained characterizations result from specifying the rules according to which the numbers are assigned, i.e., by specifying the level of measurement. He distinguishes four levels of measurement for objects o_i in a domain $D = \{o_1, \ldots, o_n\}$. Depending on the basic empirical operation involved, these measurements constitute four commonly distinguished scales: nominal, ordinal, interval and ratio scales. Their characteristic properties are summarised in Table 3 following Stevens (1946) with additions and formalisations from Schurz, 2013, pp. 99–105. Each scale type is presented in what follows and its relationship to rating scales is discussed.

Table 3: Comparison of scale type properties combining Schurz, 2013, pp. 99–105, Stevens (1946).

	Nominal	Ordinal	Interval	Ratio
mathematical function	$f: D\ to\ \{C_1, \dots, C_n\}$	$f: D\ to\ \{1, \dots, n\}$	$f: D\ to\ Num \times \{u_f\}$ with numeric value $r \in Num$ and unit of measurement u_f	
mathematical group structure	permutation group $x' = f(x)$, $f(x)$ is one-to-one substitution	isotonic group $x' = f(x)$, $f(x)$ is any monotonic increasing function	general linear group $x' = a\,x + b$	similarity group $x' = ax$
function of numbers	determine equality	determine rank order	determine equal differences	determine equal ratios
use	classification	comparison of relative position	comparison of differences	comparison of ratios
properties	$C_1, \dots, C_n$ a) mutually exclusive b) exhaustive	a) reflexivity $\forall x: x \leq_P y$ b) transitivity $\forall x, y, z: x \leq y \wedge y \leq z\ to\ x \leq z$ c) trichotomy $\forall x, y: x \leq y \vee x \geq y$		monotonicity $f(a \circ b) > f(a), f(b)$ (additivity $f(a \circ b) = f(a) + f(b)$)
associated data	frequency	rank order	meaningful interval differences	meaningful absolute differences
central tendency	mode	median	mean	geometric mean coefficient of variation
undetermined	order	intervals	absolute zero point	unit

Measurements resulting in nominal scales are those in which all objects o_i in D are assigned to classes. While there is not necessarily a fixed rule to determine how the classes are derived, Stevens identifies the rule for nominal scales as follows: “Do not assign the same numeral to different classes or different numerals to the same class.” (Stevens, 1946, p. 679), i.e., the assignment of objects to subdomains has to be definite. If numbers are used in nominal scales, their only function is to label the classes and they could be exchanged with letters without loss of information. The only numerical information derivable from a nominal scale is the frequency of objects in the different classes. Possible statistical procedures can be done based on frequency tables containing the number of objects in the different classes. Nominal scales are

constituted by classificatory concepts which divide a domain into subgroups. For an efficient scale, these subgroups should be mutually exclusive (i.e. no object from the domain belongs to more than one group) and exhaustive (each object in the domain belongs to one of the subgroups) (Schurz, 2013, p. 99).

A rank order constitutes an ordinal scale and is obtained by assigning the objects o_i in D to rank groups. Ordinal scales are constituted by comparative concepts that organise the objects into a rank order. The most common example is the "smaller-than" comparison, which orders the objects in a domain according to their size. Each object has a definite position relative to the others, which makes comparative statements about the objects from the domain, like "x is smaller than y", possible. Ranks allow the comparison of objects relative to the standard according to which they are ordered, but they do not allow a quantitative comparison between the objects, as the distances between ranks are meaningless. Data from judgement on Likert-scales is often considered to be ordinal – the distance between "strongly disagree" and "disagree" is not necessarily the same as the distance between "neutral" and "disagree". However, it makes sense to say, "I agree more strongly to statement A than to statement B" and in this sense they can be ordered. Finding one exclusive rank order for the objects in a domain requires the comparison relation to be reflexive (x is equal to x), transitive (if x is smaller than y and y is smaller than z then x is smaller than z) and connexive (everything in the domain is comparable to everything else in the domain) (cf. (Schurz, 2013, p. 101)). In his famous article "Features of similarity", Tversky (1977) argues that none of these properties holds for similarity judgments (see section 4.4). As typicality judgments are often interpreted as a special case of similarity judgement, namely similarity of a subcategory to the category prototype, his results are important for this thesis. Furthermore, typicality judgements are usually done on Likert scales which are popularly argued to result in ordinal data.

On the interval level, data is measured in equal units of measurement, like temperature in Celsius or IQ scores, which makes the differences between units meaningful, but not their ratios. On the ratio level, in which in addition to equal units, a fixed zero point is defined, like temperature in Kelvin or weight in kg, which leads to meaningful ratios. Interval and ratio scales are a function from objects o_i in a domain D to a numerical value r in a unit u_f, formally $f(o_i) = r[u_f]$, like $\text{length}(x_i) = 5\ [\text{cm}]$ (Schurz, 2013, p. 102). Ratio scales have a fixed zero point, for example the temperature in Kelvin is zero when enthalpy and entropy of an ideal gas are 0. A fixed zero point validates ratio statements like "the value of x is twice the value of y". These statements are not valid in interval scales, as the intervals are fixed by convention and internally consistent, but do not have stable ratios. 20 °C is not half as warm as 40 °C. However, the intervals are meaningful and the distance between 0 °C and 20 °C is the same as the one between 20 °C and 40 °C. (Schurz, 2013, pp. 102–103).

In relation with prototype research, nominal, ordinal and interval scales can be found. Nominal data is found in category membership questions – participants are asked whether a certain subcategory belongs to a certain category (like "are tomatoes fruit?"). Typicality is measured on scales with several gradations from 5-point- to 20-point-scales or collected as rank-order data. The most common assumption for typicality, along with many other scales measuring psychological constructs, is that they lie on an interval or quasi-interval (Coolican, 2009, p. 254) scale, which is implicit when authors report means and standard deviations for typicality ratings.

Figure 8 compares ratio scales, interval scales and typicality rated on a 7-point scale and their conversions to ordinal and nominal scales. It illustrates that it is possible to transform measurements from a higher level to a lower level, for the example of weight measurements in kilogram, body temperature measurements in Celsius and typicality ratings. As both interval and ratio scales provide numerical values of a quantity, transforming them to an ordinal scale is simply a matter of ordering the measured values according to the standard of comparison. The transformation to a (quasi-)nominal[4] scale is more subjective as it requires a decision as to which values can be meaningfully grouped together and how to label these groups. This process becomes important when we classify things as highly, medium or lowly typical. These groupings should be understood as referring to certain areas on the scale and not as precisely determined points.

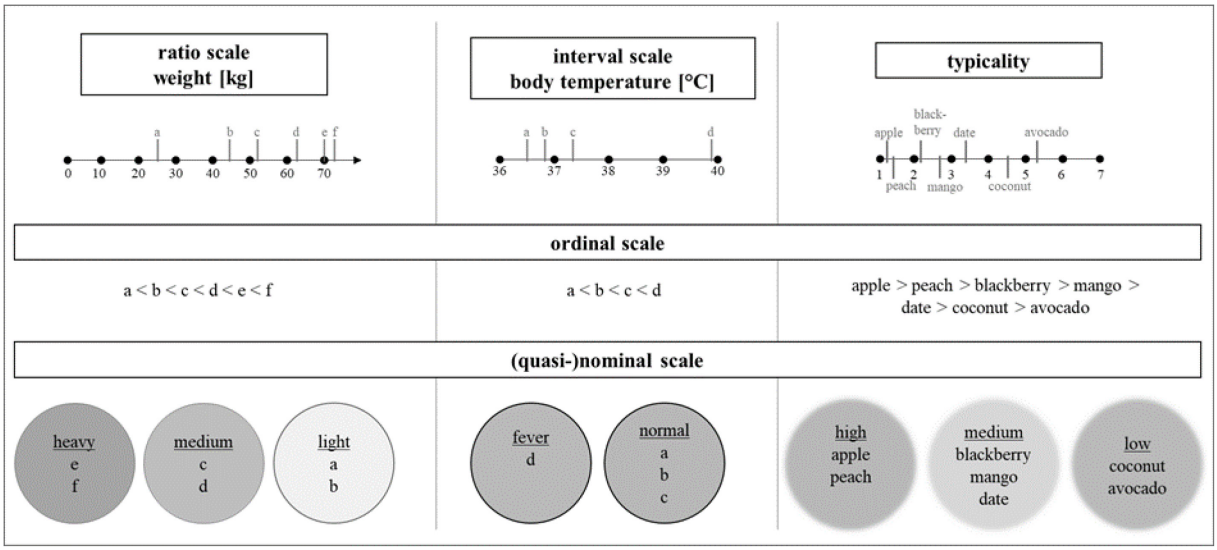

Figure 8: Examples for ratio scale, interval scale and typicality scale.

[4] It is quasi-nominal because the values are still ordered implicitly.

3.1.2 **Central tendency**

Typicality orderings are usually presented as an ordered list of the mean values of several participants' ratings. The different descriptive statistics will be presented with examples for *apple*, *fig* and *coconut* typicality ratings made on a 7-point scale by 30 participants (for details see section 6.3.4) which are in Table 4.

Table 4: Typicality ratings for apple, fig and coconut by 30 participants with mean, median, SD, IQR, and the first two modes. Ratings were done on a scale from 1 (very good example) to 7 (very bad example).

participant	apple	fig	coconut
S1	1	4	5
S2	1	7	7
S3	1	3	2
S4	1	1	7
S5	1	5	6
S6	1	1	4
S7	1	1	1
S8	1	4	7
S9	1	1	2
S10	1	4	7
S11	7	7	4
S12	4	7	7
S13	1	4	4
S14	2	4	6
S15	1	6	3
S16	1	7	7
S17	1	3	6
S18	2	4	2
S19	1	1	2
S20	1	4	6
S21	1	1	7
S22	1	1	5
S23	1	1	3
S24	1	3	3
S25	1	5	4
S26	1	5	7
S27	1	1	4
S28	1	4	4
S29	1	4	4
S30	1	6	4
mean	1.4	3.6	4.7
SD	1.2	2.1	1.9
median	1.0	4.0	4.0
IQR	0.0	4.0	4.0
mode 1	1	4	7
mode 2	1	1	4

Raw data can be visualised in scatter plots, which represent each observation with a dot. The x-axis usually stands for the independent variable, or the input, and the y-axis for the dependent variable, or the output. For the typicality ratings, the independent variables are the subcategories, and the dependent variable is their typicality rating. If values occur multiple times, it is useful to add some random noise to the data by multiplying the data with a uniform random variable in a specified interval. This process is called jittering and the specified intervals are the jitter width and height. The interval in which noise is added should be smaller than the distance between the rating points, then it is visible to which rating each point belongs. If the distance between ratings is 1, which is common for typicality ratings, a jitter interval of 0.2 is ideal to visualise all ratings while keeping the possibility to identify to which rating each dot corresponds, all 1 ratings are for example found in the interval [0.8,1.2]. Figure 9 shows the typicality ratings for *apple*, *coconut* and *fig* in a scatterplot where the typicality scale is plotted on the y-axis against the subcategories, with a) no jitter, b) jitter = 0.2, c) jitter = 1. When no jitter is applied like in Figure 9a), only qualitative information about the ratings is visible on the plot: *apple* received the ratings 1, 2, 4 and 7, *coconut* received ratings on the whole scale and *fig* received all ratings except 2, but the plot does not show how many of each. With an accurate jitter like in Figure 9b), it is visible that most ratings for *apple* were 1, only 2 ratings were 2 and 4 and 7 were chosen once. For *coconut*, 2, 4 and 7 were chosen more than 3 times, the other 4 ratings 3 times or less. For *fig*, about one third of ratings were 1 and the other two thirds spread between 3 and 7. With a jitter as high as the distance between two rating points like in Figure 9c), it is no longer possible to differentiate between the ratings.

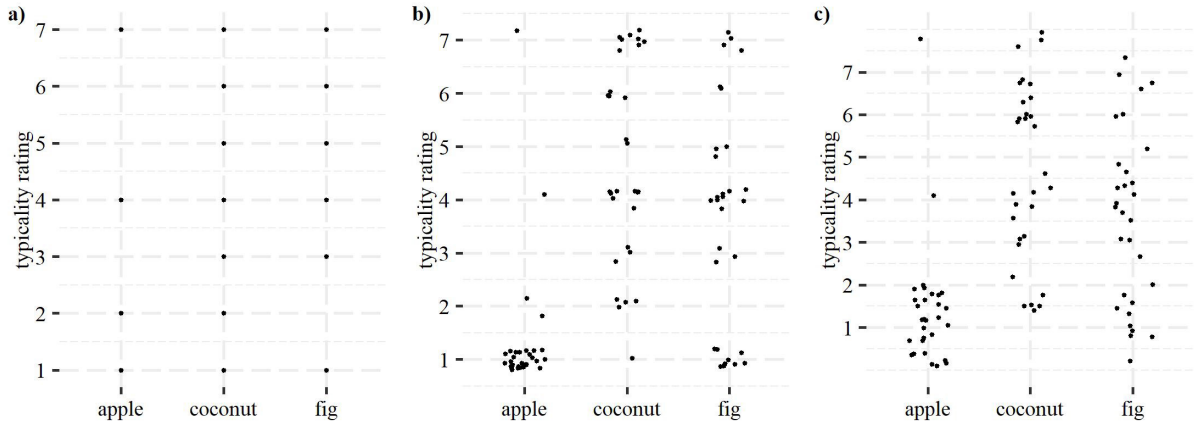

Figure 9: Scatter plot of apple and fig typicality ratings with 3 different jitters: a) no jitter, b) jitter=0.2, c) jitter=1.

Lists and plots of raw data are more difficult to evaluate the larger they get. It is therefore useful to summarise the raw data with the goal to accurately describe them with few parameters that reflect their central tendency, dispersion and the shape of their distribution.

The first important question is where the central tendency of the data lies. It can be calculated as the mean, median or mode. The mode is the most frequent value occurring in the dataset and requires a frequency table that contains the number of observations for each possible value. If two or more values occur equally often or constitute a local maximum, the dataset has two or more modes, and the distribution is called bi- or multimodal. The mode is the only measure of central tendency that can be applied to nominal data because it does not assume an underlying order. It only reflects the dominant response(s). But it is also informative for data on ordinal, interval and ratio scales. The mode is, contrary to median and mean, a value that corresponds

to a datapoint that is really occurring in the data. As typicality is assumed to be an intersubjectively stable variable, typicality ratings should have a clear modal value in which most, if not all, observations are found. The opposite of this would be a uniform distribution, where each value occurs equally often. Uniform distributions are observed for random processes, for example for sufficiently often repeated coin tosses. If typicality ratings had a uniform distribution, typicality would not be an interesting cognitive variable because it would not be intersubjective. A multimodal distribution of typicality ratings would show that there is no agreement between all participants, but between subgroups. This could be due to the experimental design being differently interpreted by different participants. Another explanation is that there are different points of view that lead to multi-modal ratings. For example, one could say that a fig is a typical fruit because it is clearly a fruit, or one could say it is medium typical because it is a fruit, but it does not share many properties with other fruit. As typicality ratings are assumed to be made intuitively, this could mean that there are different ways in which people view categories, classical (each fruit is a typical fruit) or graded.

Table 5 shows the frequency table for the typicality data from Table 4, from which the mode can be identified. The modes are 1 for *apple*, 1 and 4 for *fig* and 4 and 7 for *coconut*.

Table 5: Frequency table for the typicality rating for apple and fig.

	1	2	3	4	5	6	7
apple	26	2	0	1	0	0	1
fig	9	0	3	9	3	2	4
coconut	1	4	3	8	2	4	8

Frequency counts can be visualised in histograms, which show the distribution of ratings on the scale by counting how many ratings lie in each specified interval, called bin. As each observation is treated as one count in a category of values with a specific ordering, interval- or ratio-level data can be transformed into ordinal data depending on the width of the bins. The minimal bin width is the minimal distance between ratings, for the typicality ratings this is 1. Figure 10 shows histograms of the data, a) for 7 bins and b) for 2 bins. With one bin for each rating, the histogram represents the frequency table, and the modes can be identified. A larger bin width makes the data more compact: if the data is split into two intervals from 1 to 3 and 4 to 7, the information is more condensed and the qualitative information becomes easier accessible: *apple* received a lot more typicality ratings at the upper end of the scale, while for *coconut* most were at the lower end of the scale. For *fig*, the ratings are almost evenly distributed between the higher and the lower end.

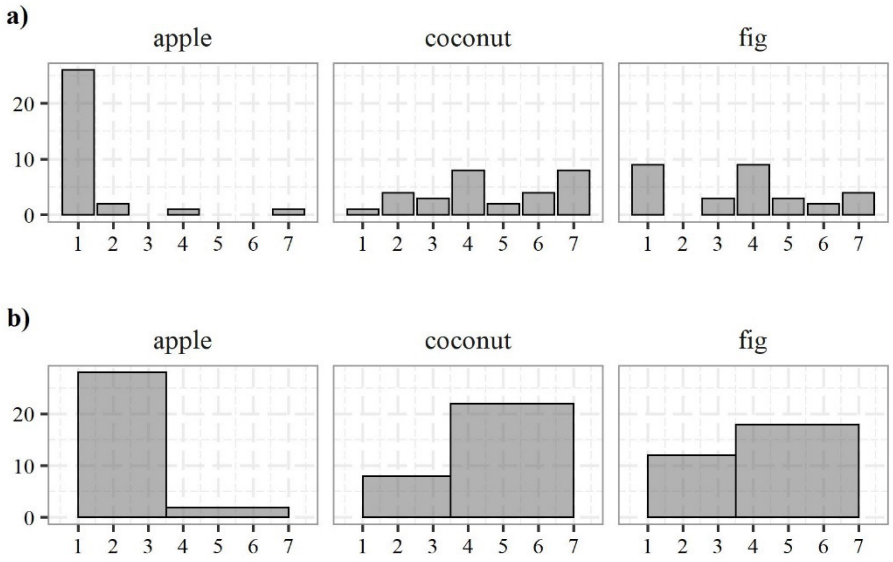

Figure 10: Histograms for apple, fig and coconut typicality ratings a) 7 bins, b) 2 bins (right-closed).

The median is the central value of a dataset and divides it into two equally sized sets, one of which contains all the values below the median and the other all the values above. It is obtained by ordering the values in the dataset and then identifying the value in the middle. For an odd number of observations, the median is exactly the value dividing the ordered data into two halves and for even numbers it is the mean of the two values in the middle. The typicality data ordered by size are in Table 6. The medians are 1 for *apple* and 4 for *fig* and *coconut*. While the median is representative of the voting behaviour for *apple*, it does not reflect the central tendency of *fig* and *coconut* ratings.

Table 6: Ordered ratings for apple, fig and coconut typicality.

apple	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	4	7
fig	1	1	1	1	1	1	1	1	1	3	3	3	4	4	4	4	4	4	4	4	5	5	5	6	6	7	7	7	7
coconut	1	2	2	2	2	3	3	3	4	4	4	4	4	4	4	4	5	5	6	6	6	6	7	7	7	7	7	7	7

The median is robust to outliers, because only the centre of the values in the dataset is used for its calculation. This can, in small datasets or in datasets with large value differences, also lead to a misrepresentation of the dataset: the median of 5, 6, 8, 20, and 50 is 8. The distances between values are not reflected in the median. If outliers are expected in the experimental setting and have no influence on the conclusion, like when participants sometimes take longer in a reaction time experiment for external reasons like drinking a glass of water, the median reflects the central tendency better because it neglects them.

The measure of variance for the median is the interquartile range (IQR). It is calculated by splitting the values into quartiles and subtracting the median of the upper one (third quartile, Q_3) from the median of the lower one (first quartile, Q_1). It corresponds to the difference between the medium 50% of datapoints. For even data sets with 2n datapoints and odd data sets with 2n+1 datapoints, the first quartile is the median of the n smallest data points and Q_3 is the median of the n largest data points. Outliers are usually considered to be those values which deviate more than 1.5 IQR from the lower or upper quartile.

The 3 quartiles for the typicality data are in Table 7. The IQR of the typicality ratings is 0 for *apple* and 4 for *fig* and *coconut*. Combined with the median this reflects the ratings for *apple* quite well: the centre of the ratings is 1 and the medium 50 % also lie in 1. *Fig* and *coconut* have a very high IQR, reflecting that the medium 50 % of ratings are spread on almost the whole scale. Information that is lost in the IQR is the direction of the dispersion: note that the upper quartile for *fig* has a median of 5 and is thus only 1 scale point away from the median, while the lower one has a median of 1 and is 3 points away from the median.

Table 7: Ordered ratings from Table 4 with indication of first, second and third quartile.

							Q1							Q2								Q3						
apple	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	4	7
fig	1	1	1	1	1	1	1	1	1	3	3	3	4	4	4	4	4	4	4	4	5	5	5	6	6	7	7	7	7
coconut	1	2	2	2	2	3	3	3	4	4	4	4	4	4	4	4	5	5	6	6	6	7	7	7	7	7	7	7	7
							3															7						

IQR = Q3 - Q1 = 4

Medians are usually visualised in boxplots. Boxplots for the typicality data are in Figure 11. The median is represented as the line in the box and the lower and upper end of the box represent the lower and upper quartile. The medium 50% of the data lie inside this box. The length of the box thus corresponds to the IQR. The lines that come out of the box are called whiskers. They end at the minimum and maximum, so between the borders of the box and the ends of the whiskers are the lower and upper 25% of the data, except for outliers. Outliers are values which deviate more than 1.5 IQR from the lower or upper quartile and are represented as points. To get a more complete picture of the data, the individual ratings can be overlayed. The boxplot for *apple* shows that the median of the ratings is 1 (black line) and the fact that there is no box around the median line means that the IQR is 0, the medium 50% of the data are exactly 1. It has outliers on 3 values: at 2, at 4 and at 7. Thanks to jittering the outliers, it is visible that there are 2 outliers at rating 2. The boxplot for *fig* shows that the whole scale was used by the participants. The plot starts at 1 and the whiskers end at 7. The IQR is 4 and the ends of the box show that the medium 50% of data lie between 1 and 5. However, the box is much longer below the median than above, which shows that 25% of participants chose a rating between 4 and 5, while the other 25% chose a rating between 4 and 1. Thus, from the information in boxplots, statements about the symmetry of the rating distribution are possible. The information that there is a second mode at 1 is however lost. A very similar picture can be seen in the boxplots for *coconut*.

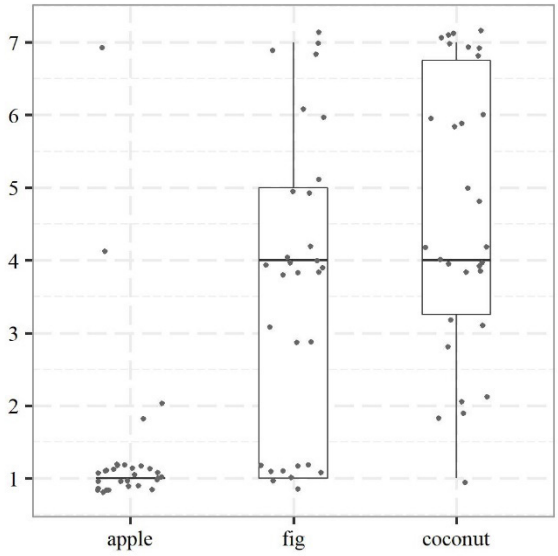

Figure 11: Apple, fig and coconut typicality ratings (grey jittered points, width and height = 0.2) and summarised in boxplots.

The mean is calculated by taking the sum of all ratings x_i and dividing it by the number of ratings N:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{N}.$$

The means of the typicality ratings are 1.4 for *apple*, 3.6 for *fig* and 4.7 for *coconut*. The mean is the basis of parametric statistical methods and accounts for the distances of the collected values, which the median does not. Its sensitivity can be seen as a disadvantage because it makes the mean easily distorted by outliers, which increase or decrease it drastically. This can result in a misrepresentation of central tendency. Figure 12 visualises the mean and median for the typicality ratings in histograms as straight resp. dotted line.

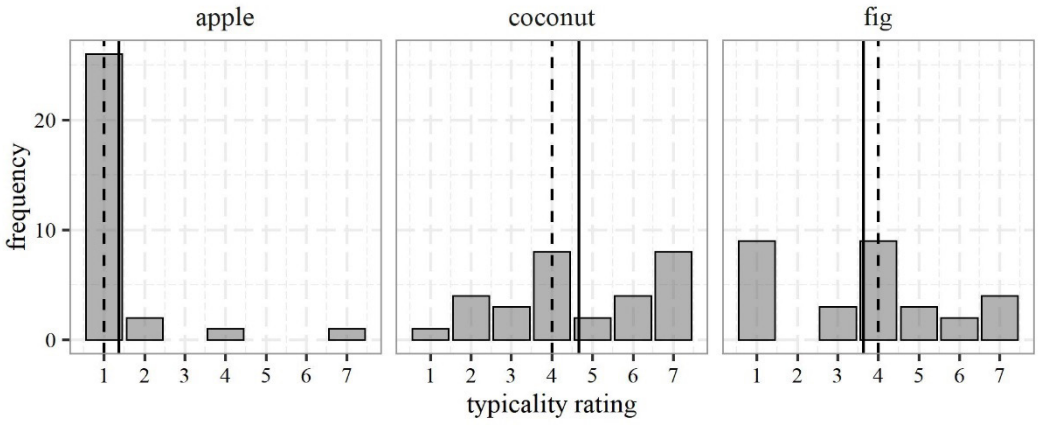

Figure 12: Histograms for apple, fig and coconut typicality ratings with vertical lines showing the mean (straight line) and median (dotted line).

It is important for the interpretation of means to examine the dispersion of the data. Most commonly reported together with the mean is the standard deviation (SD). It measures the average of the deviations from the mean, calculated by taking the square root of the sum of

squared deviations of each datapoint $x_1, \ldots, x_N$ from the mean $\bar{x}$, also called the sum of squares of x (SS_x), divided by $N - 1$:

$$SD = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{N - 1}}.$$

The denominator is N-1 because the average squared deviation SS in the nominator is based on the mean of the sample but the SD is supposed to estimate the population variance. What is searched is the average squared deviation of the data from the actual population mean, which is unknown. As the data is always closer to its own mean than to the population mean, the estimation $\frac{SS}{N}$ is an underestimate. To correct for this, the variance of the sample mean around the population mean ($\frac{V}{N}$) is added in:

$$\text{estimated } V = \frac{SS}{N} + \frac{V}{N}$$

$$\text{from which } NV = SS + V \Leftrightarrow (N - 1)V = SS \Leftrightarrow V = \frac{SS}{N - 1}$$

$$SD = \sqrt{V} = \sqrt{\frac{SS}{N - 1}}$$

The correction has a substantial influence on the SD of data with small samples, while the influence is negligible with large samples. If data on the whole population is available, the denominator is N. SD^2 is often referred to as variance v. In a normal distribution, 68.26% of data lie within the interval [$\bar{x}$-SD, $\bar{x}$+SD], 95.44% within [$\bar{x}$-2SD, $\bar{x}$+2SD] and 99.74 within [$\bar{x}$-3SD, $\bar{x}$+3SD]. It is a rule of thumb that for data with approximately normal distribution, two thirds of datapoints are within one SD from the mean. Small SDs indicate small deviations from the mean and thus a good reflection of the central tendency, while a high SD can mean a bad reflection of central tendency or be due to the influence of a small number of extreme outliers. Because all differences are squared, outliers have a stronger influence on the SD than on the mean.

The SDs for the typicality ratings are 1.2 for *apple*, 2.1 for *fig* and 1.9 for *coconut*. Despite 26 of the 30 ratings (87%) for *apple* being 1, the SD is high. This illustrates the mean's sensitivity for outliers. If the one 7-rating is removed, the SD decreases to 0.6. Visualising means is usually done by plotting the mean as points and adding error bars as whiskers which represent the SD, like in Figure 13 for the typicality data. For *fig* and *coconut*, both mean with SD and median with IQR reflect the central tendency poorly, if at all, as they show dispersion on almost the whole scale. As the SD is symmetric around the mean, it gives no information on the position of the outliers.

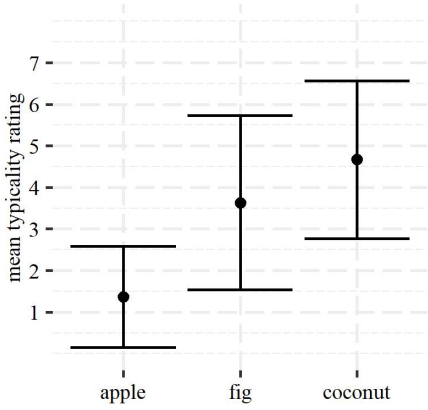

Figure 13: Apple, fig and coconut mean typicality ratings in scatterplots with error bars.

All three standard measures have advantages and disadvantages. If the datapoints have a symmetric unimodal distribution, mean, median and mode have the same value. We have seen that the information on the bimodal distribution of *fig* and *coconut* typicality ratings is neither reflected in the mean nor in the median. If multimodality is not ruled out by a low SD or IQR, it is advisable to consult frequency tables or to apply the dip test presented in Hartigan and Hartigan (1985), which determines multimodality "by the maximum difference, over all sample points, between the empirical distribution function, and the unimodal distribution function that minimizes that maximum difference" (p. 70). Its results are the dip test statistic which is the "maximum distance between the empirical distribution and the best fitting unimodal distribution" (p. 80) and a p-value. A distribution is considered multimodal if the p-value is ≤.05. Because it is designed for continuous data (ratio or interval scales), its use for discrete data (on ordinal scales) is restricted, as the dip test is sensitive for local maxima if they are very close on the scale. This is too strong a criterion for typicality scales, because ratings right next to each other do not seem to reflect noteworthy differences of participants' opinion. To account for this, the analysis of the dip p-values can be combined with an analysis of the IQR. A high IQR shows that the medium 50% of ratings are spread on the scale and if the dip p-value is low in addition, there are two or more maxima, which is a good indication for a multimodal distribution.

3.1.3 **Correlation**

Correlations are important in this thesis in two cases: in the typicality meta-analysis (section 3.2) they are one of the measures used to determine how intersubjectively stable typicality ratings are and they are used to evaluate the results of the typicality predictions with different datasets in chapters 7 and 8.

Correlation coefficients describe the linear relationship, or covariance, between two variables in normalised values in the interval [-1,1], where -1 is a perfect negative correlation, 0 indicates no covariance and 1 is a perfect positive correlation. The covariance cov_{xy} of two datasets x and y is the sum of the cross-products of distances of each shared datapoint from the mean of the dataset divided by N-1:

$$\mathrm{cov}_{xy} = \frac{\sum(x - \bar{x})(y - \bar{y})}{N - 1}.$$

Covariance corresponds to the formula of the variance (SD^2) when y is replaced with x. The maximal values of cov_{xy} are $\pm SD_x SD_y$. The most common correlation coefficients are

Pearson's r for interval level data and Spearman's ρ for ordinal data. The correlation coefficient formula is the covariance divided by the product of the SDs of both datasets. This normalises the values to ± 1:

$$r = \frac{cov_{xy}}{SD_x SD_y} = \frac{\sum(x - \bar{x})(y - \bar{y})}{(N-1)SD_x SD_y}.$$

For Spearman's rank order correlation coefficient ρ, the datapoints of x and y are rank ordered and the ranks r_x and r_y are the input for the formula.

$$r = \frac{cov_{r_x r_y}}{SD_{r_x} SD_{r_y}} = \frac{\sum(r_x - \bar{r_x})(r_y - \bar{r_y})}{(N-1)SD_{r_x} SD_{r_y}}$$

According to Cohen (1988), a strong correlation, i.e., a correlation with a high effect size, is found with $r \geq .5$, moderate with $r \geq .3$ and weak with $r \geq .1$. Correlation coefficients are associated with p-values that reflect their significance and can be looked up based on the results of a t-test with N-2 degrees of freedom:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}.$$

Correlations are commonly taken to be significant if the associated p-value is ≤.05.

Figure 14 shows examples of correlations of normalised mean typicality ratings from different studies. Correlations are usually visualised in scatterplots (right side of Figure 13). They are an important tool in the analysis of relationships between variables because they visualise them. The regression line is the best-fitted line between the datapoints. For perfect correlations, it is a straight diagonal line on which all datapoints lie. It is interesting to compare the values for each datapoint (left side of Figure 13). The correlation between the fruit typicality data from Rosch (1975b) and Uyeda and Mandler (1980) is almost perfect with both Pearson and Spearman coefficients >.9 and it is significant ($p<.001$). An inspection of the left side shows that the means do not have the exact same values. All that matters for the correlation is that the distance from the mean value is in the same direction and has the same size. The correlation between the fruit typicality data from Rosch (1975b) and Moreno-Martínez et al. (2014) is less perfect but still high ($r = .67$, $\rho = .8$) and significant ($p \leq .001$). The points lie on a less perfect line. The correlation between the vegetable typicality data from Uyeda and Mandler (1980) and Moreno-Martínez et al. (2014) has a small effect size ($r = .26$) and is not significant ($p>.05$). The points resemble a cloud and do not show a linear relationship, the regression line is almost flat.

If the correlation coefficient is low or the linear regression has no satisfying results, this does not mean that there is no relationship between the two variables. The best-fitted line between the datapoints is not straight when the relationship is not linear. Other functions could be better suited to describe the relationship, for example curved functions like binomial or exponential functions.

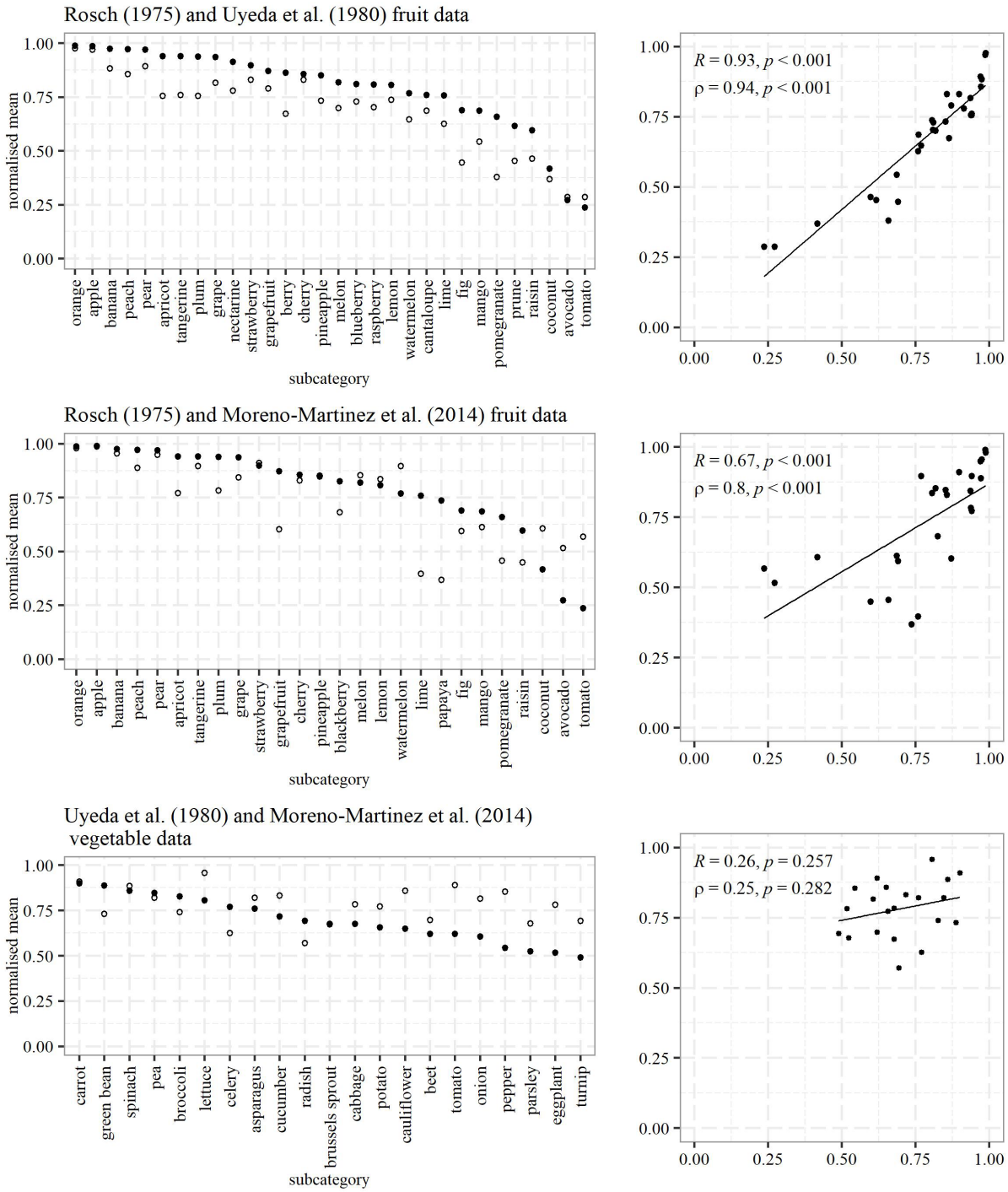

Figure 14: Examples for correlations from the typicality meta-analysis.

3.1.4 **Reliability**

When data is collected to draw inferences from the participants' replies to the whole population, the data has to be reliable, i.e., it should not exhibit a high intersubjective variance. This kind of reliability is more precisely called "inter-rater reliability" or "intersubjective reliability" (as opposed to other kinds of reliabilities).

Reliability is usually determined with split-half correlations, obtained by creating two equal groups of participants, determining the mean for each half and then calculate the correlation between these two means. In De Deyne et al. (2008) this was done for example for 10,000 randomizations.

Cronbach's alpha is often used to assess the reliability of scales and is a function of the average correlation $\bar{r}$ between the scale items and the number of scale items N:

$$\alpha = \frac{N\,\bar{r}}{1 + (N-1)\bar{r}}.$$

α corresponds to the mean of all possible split-half correlations. It can be performed on the items of a questionnaire over all participants, but also for the participants over all items, to get an idea in how far the participants' judgments differ from the mean without them. The latter was used in this thesis to spot participants as outliers who do not reflect the common ground. The statistic's output are the coefficient alpha, which indicates a reliable scale for values >.75 (Coolican, 2009, p. 195), and the correlations of each participant or scale item with the mean of the remainder.

A different kind of reliability is the reliability of means. It can be assessed by computing confidence intervals. The guiding question is in how far the observed mean (sample mean) reflects the population mean or how confident we can be that the mean we observed reflects the "true" mean that would result from complete observation. It is based on the standard error (SE), the root of the squared sample SD divided by the number of datapoints N:

$$SE = \sqrt{\frac{SD^2}{N}}.$$

The SE is an estimate of the SD of a large sampling distribution of the means of N datapoints. It gets smaller with a higher sample size and with a lower observed SD. The confidence interval is defined as $[\bar{x} - t \cdot SE, \bar{x} + t \cdot SE]$, where t depends on the level of confidence required. For the commonly used 95% confidence with 30 datapoints it amounts to 2.042, for 98% confidence it is 2.457. These numbers are derived from Student's t distribution, which closely approaches the normal distribution as the sample size N increases. 95% confidence means that the real mean will lie in the confidence interval with 95% certainty or that the chance with which the confidence interval contains the mean is 95% or that if the sampling would be repeated 100 times, the mean would be in the confidence interval in 95 of these times.

A confidence interval can also reflect the reliability of correlations. To determine the mean correlation and the SE, the jackknife technique can be used, as described in chapter 11 in Efron and Tibshirani (1993). Here, all N observations used for the correlation are systematically left out and the new correlation coefficient for N-1 observations is computed, which results in N correlation coefficients r_i:

$$r_i = \frac{1}{N-1} \sum_{j=1,J\neq i}^{N} x_j \, , i = 1, \ldots, N.$$

The mean of the n correlations is $\bar{r}$. The SE of the correlation can then be determined as:

$$SE_{jack} = \sqrt{\frac{N-1}{N} \sum_{i=1}^{n} (r_i - \bar{r})^2}.$$

Table 8 illustrates the steps of the jackknife procedure exemplarily for the correlations of the fruit typicality ratings reported in Rosch (1975b) and Moreno-Martínez et al. (2014). The r_i for the different SCs give an indication on which pairs (de-)stabilise the correlation: leaving out *apple*, which received the same rating in both datasets, reduces the correlation from .67 to .66 and leaving out *papaya*, which received a high rating in Rosch's data (.74) and a low rating in Moreno-Martinez et al.'s data (.37), raises it to .71. The SE determined according to the formula above is .07 and the correlation has a 95% confidence interval of [.53,.82].

Table 8: The jackknife procedure for typicality data.

subcategory	Rosch 1975	Moreno-Martinez et al. 2014		r_i
apple	.99	.99		.66
apricot	.94	.77		.68
avocado	.27	.52		.66
banana	.98	.96		.66
blackberry	.83	.68		.68
cherry	.86	.83		.67
coconut	.42	.61		.68
fig	.69	.59		.67
grape	.94	.84		.67
grapefruit	.87	.60		.69
lemon	.81	.84		.68
lime	.76	.40		.71
mango	.69	.61		.67
melon	.82	.85		.68
orange	.99	.98		.66
papaya	.74	.37		.71
peach	.97	.89		.66
pear	.97	.95		.66
pineapple	.85	.85		.67
plum	.94	.78		.67
pomegranate	.66	.46		.67
raisin	.60	.45		.66
strawberry	.90	.91		.67
tangerine	.94	.90		.67
tomato	.24	.57		.70
watermelon	.77	.90		.69
	r	.67	$\bar{r}$	.67

3.2 A meta-analysis of typicality experiments

The main observation that led to the development of prototype theory is that typicality ratings are intersubjectively stable and are therefore assumed to provide insights into the internal structure of category representations. It is the reason why typicality is assumed to play a fundamental role in cognition. Eleanor Rosch makes this clear in the following statement:

> "It is by now a well-documented finding that participants overwhelmingly agree in their judgements of how good an example or clear a case members are of a category […]. Were such agreement and reliability in judgment not to have been obtained, there would be no further point in discussion or investigation of the issue." (Rosch, 1978, p. 11)

Rosch quotes four studies to substantiate the claim of the high intersubjective agreement of typicality ratings, which was a lot in the 1970s when typicality research began. Today, many more studies are available, and it is an interesting starting point for my investigation to analyse how stable these ratings are in the lights of a considerable amount of new evidence. The availability of data in different languages (English, German, Dutch, Spanish, French) and from different decades (1970s to now) allows a contrasted examination of their intersubjective stability and can shed light on the question in how far concept formation depends on external factors like the living environment and which category members are encountered in it.

In my meta-analysis I compare typicality data for fruit and vegetables reported in Rosch (1975b), McCloskey and Glucksberg (1978), Uyeda and Mandler (1980), Malt and Smith (1982), Hampton and Gardiner (1983), Barsalou (1985), Brown and Semrau (1986), Schwanenflugel and Rey (1986), Smith et al. (1988), Ruts et al. (2004), De Deyne et al. (2008), Schröder et al. (2012), Moreno-Martínez et al. (2014), the data collected in the course of this thesis (section 6.3.4) and an unpublished set of typicality ratings by French speakers, kindly provided by James Hampton.

The studies differ in experimental design questions, specifically in how categories and subcategories were chosen (chapter 3.2.1), the typicality scales used to collect the ratings (3.2.2), the wording of the instructions (section 3.2.3) and the choice of participants (3.2.4). In section 3.2.5, I compare all datasets and in section 3.2.6 I analyse the correlation matrices. In section 3.2.7, the rating behaviour of participants is analysed for the four datasets for which all individual ratings are available.

3.2.1 Choice of categories and subcategories

The choice of categories for typicality experiments is usually made to represent a wide variety of concepts, which are generally known and occur frequently in everyday life. One study is often used as basis for the decision which categories to include: productive frequency data from Battig and Montague (1969). Battig and Montague (1969) made a large investigation of 56 categories. They asked 442 participants to write down all words associated with the categories they could think of in 30 seconds and report frequency tables of all responses. The number of times an associate was generated is called its productive frequency (PF). For example, *dog* was generated by 426 participants for the category *a four-footed animal*, this means that the PF of *dog* for *four-footed animal* is 426. Selected categories from those that were used in their study used are used in typicality studies.

Table 9 shows all categories used in at least 3 of the 15 datasets collected for the meta-analysis. *Fruit, vehicle, vegetable, bird, clothing, furniture* and *sport* were used in 10 or more of the datasets. The usual prototype categories are supercategories of basic level categories. They are common in everyday life and everyone can be assumed to know multiple examples, as

demonstrated in Battig and Montague (1969). Schurz (2012) proposes to restrict the domain of prototype concepts to those categories that were shaped by cultural or biological evolution and therefore are describable with a prototypical norm state in which most members are most of the time (see chapter 5). This is the case for all categories in Table 9.

Table 9: Categories used in at least 3 studies collected for the meta-analysis.

category	number of datasets	Rosch 1975	McCloskey et al. 1978	Uyeda et al. 1980	Hampton et al. 1983	Malt et al. 1982	Malt et al. 1984	Barsalou 1985	Brown et al. 1986	Schwanenflugel et al. 1986	Smith et al. 1988	Ruts et al. 2004	De Deyne et al. 2008	Schröder et al. 2012	Moreno-Martinez et al. 2014	Hampton French data
fruit	15	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
vehicle	13	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	
vegetable	12	✓	✓	✓	✓			✓	✓		✓	✓	✓	✓	✓	✓
bird	12	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓		
clothing	12	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	
furniture	12	✓	✓	✓	✓	✓	✓	✓	✓	✓				✓	✓	✓
sport	10	✓	✓	✓	✓			✓	✓			✓	✓	✓		✓
musical instrument	7			✓					✓	✓		✓	✓	✓	✓	
weapon	7	✓		✓	✓			✓	✓	✓			✓			
tools	7							✓	✓			✓	✓	✓	✓	✓
fish	7		✓		✓	✓	✓					✓	✓			✓
insect	6		✓		✓							✓	✓		✓	✓
flower	5				✓	✓	✓		✓						✓	
animal	4		✓						✓					✓	✓	
kitchen utensil	4		✓	✓									✓		✓	
profession / occupation	4								✓			✓	✓	✓		
toy	4	✓		✓					✓		✓					
carpenters's tool	3	✓	✓	✓												
mammal	3								✓			✓	✓			
part of the human body	3			✓						✓					✓	
tree	3			✓	✓										✓	

A trend in the selection of SCs is to base it on PF data, either collected within the typicality study or selected from Battig and Montague (1969). The criteria for inclusion differ. Rosch (1975b) used all SCs with PF>10 in Battig and Montague (1969) and chose additional ones that were produced by fewer participants to have 50-60 possible category members per category. Uyeda and Mandler (1980) chose the first 30 SCs from Battig and Montague (1969), thus only SCs with a higher PF. Hampton and Gardiner chose 34 to 55 SCs from Battig and Montague (1969) and Rosch (1975b) and in addition included SCs for which they needed normative data. They excluded SCs that are not category members and aimed at covering the whole range of the typicality scale. Schwanenflugel and Rey (1986) chose SCs from the ones used in Uyeda

and Mandler (1980) which have precise semantic equivalents in English and Spanish. They then added subcategories "to make the norms more extensively describe the range of exemplars familiar to South Florida residents" (p.152).

In some studies, PF data was collected before typicality ratings. Barsalou (1985) included all SCs generated by more than 1 out of 38 participants in 15 seconds, in which resulted in 9 to 24 SCs (median = 19.83), with 19 for fruit and 21 for vegetables (p. 634). Moreno-Martínez et al. (2014) included all SCs that were named in 60 seconds in a former study, excluding repetitions and "intrusions (e.g. generating *pear* in the category of *trees*)" (p.1090). They report results for 41 SCs for both fruit and vegetables. Schröder et al. (2012) collected PF data from 20 participants, instructed to "write down as many examples as they could think of for each semantic category" (p. 383) with no time restriction. They excluded SCs that were judged by 2 independent judges not to belong into the category, homographs (like kiwi as a bird and fruit) and merged synonyms. The resulting list of 1,123 SCs for 11 categories was used for the typicality ratings, but only results for those are reported which were not rated to be unfamiliar or not a category member by at least 25% of the participants and had SDs smaller than 2. Ruts et al. (2004) made a PF study as well and chose SCs intuitively from the ones generated by 120 participants which they felt would range from very atypical to very typical, aiming for 30 SCs per category, which was not possible in all cases due to some categories being not familiar enough (for *amphibians*, only 3 biologically correct SCs were produced). De Deyne et al. (2008) chose the same SCs and added for 3 additional categories a "representative sample" from Storms (2001), who replicated the procedure from Battig and Montague for the Flemish language.

Two studies do not mention an external source for the SCs they selected. McCloskey and Glucksberg (1978) used 30 SCs which they describe to range from very typical to atypical (e.g. dandelion for vegetables) to completely unrelated (e.g. steak for vegetables). Brown and Semrau (1986) chose 45 or 60 SCs per category. They state that they included "extremely atypical members (and some non-members) to provide anchor points at the low end of the typicality scale" (p. 125).

The most important difference between the criteria for SC selection is how the researchers treat the lower end of the typicality scale. While some studies focus on SCs with a high PF, others explicitly include atypical or unrelated ones. This could have an influence on the resulting typicality rating due to the range effect (e.g. Hutchinson, 1983). The range effect describes the phenomenon that the range of alternatives alters participants' judgement of stimuli.

When participants are confronted with many bad examples for a category, the good examples might seem better and receive a higher rating compared to the bad examples than when there is an equal amount of good and bad examples. In many datasets, *olive* is among the least typical SCs for *fruit*, it has for example a mean of 6.21 out of 7 in Rosch (1975b). In McCloskey and Glucksberg (1978), it has in comparison a higher mean of 4.04 out of 10. In their dataset, the least typical *fruit* SCs included are *chicken*, *carrot* and *sunflower seeds*, which are clear non-members. If *chicken* is in the list, participants rate the typicality of *olive* higher than when the least typical SC is *squash*. The question whether clear non-members should be included thus has an effect on how participants use the scale and can influence the correlations between data sets, which was also noted by Brown and Semrau (1986), who report lower correlations of their data with Uyeda and Mandler (1980) who had only included relatively common SCs in their experiment – "since their choice of instances is biased towards the typical, the range of

typicality values in their sample is restricted when compared with Rosch's norms or ours." (p. 128). Hampton (2017) also points out that the inclusion of clear non-members changes the meaning of the typicality scale: "to say that X is a typical (or atypical) Y carries the presupposition that X is indeed a Y. […] Rosch (1975) also chose to ignore this refinement […] and so typicality also has an extended meaning corresponding to something like 'typicality if it is a member and closeness to the category if it is not'" (p. 109).

3.2.2 **Typicality scales**

Like the selection of SCs, the scale design differs between the studies. There are different wordings of the scale anchors (whether they ask the participants to rate typicality or goodness-of-example), different amounts of scale points, different directions of the scale, different interpretations of the lowest scale point, and some include opt-outs for unknown or unfamiliar SCs or non-membership. The variants are summarised in Table 10. Of the 14 studies analysed, 10 use the goodness-of-example wording. In 9 studies, the highest rating corresponds to the highest level of typicality. Half of the studies include non-membership as the lowest scale point and 2 include an opt-out for non-membership. 6 studies include an opt-out for SCs that are unfamiliar or unknown to the participant. 6 studies use a 7-point-scale, 3 a 20-point-scale, 2 a five-point-scale and the remaining 3 studies have between 8 and 11 scale points. There is a 16-year-gap between the studies: 9 were carried out between 1975 and 1988 and 5 in or after 2004.

Rosch (1975b) chose a 7-point-scale in which the lowest scale type designates both low typicality and non-membership:

> "A 1 means that you feel the member is a very good example of your idea of what the category is. A 7 means you feel the member fits very poorly with your idea or image of the category (or is not a member at all). A 4 means you feel the member fits moderately well." (Rosch, 1975b, p. 198)

Uyeda and Mandler (1980) and Brown and Semrau (1986) used the same instructions and Schwanenflugel and Rey (1986) and Malt and Smith (1984) used variants, with reversed scale order. Schröder et al. (2012) translated them and added opt-outs for unknown SCs and non-membership. Moreno-Martínez et al. (2014) also used a goodness-of-example formulation but a high-to-low 5-point-scale. In their scale, the lowest rating designates non-membership, unfamiliarity and low typicality. McCloskey and Glucksberg (1978) used a scale from 1 to 10 (1 = extremely atypical (i.e., unrelated), 10 = very typical) and gave the option to mark the word as unknown. Barsalou (1985) used a 9-point-scale (1 = poor example, 9 = excellent example). Ruts et al. (2004) used a 20-point-scale (1 = very atypical or unrelated, 20 = very typical). De Deyne et al. (2008) collected ratings on the same scale and also goodness-of-example data (1 = very bad example, 20 = very good example). Hampton and Gardiner (1983) modified former instructions by allowing the reply that the object in question is not in the category and to leave a blank when they didn't know the word. Their scale ranges from 1 to 5 (1 = represents a very typical instance of a category, 5 = represents a very atypical instance) and additionally 6 means that the object in question is not in the category at all. They included this number in the calculation of means to achieve a more fine-grained analysis but found a correlation of .94 to the alternative of omitting it (p.495).

Table 10: Typicality scale anchors and opt-outs found in the literature, sorted by wording, scale direction and treatment of category membership (contained in lowest scale point or as an opt-out or not contained at all).

wording	scale direction	membership	Article	Typicality scale anchors: High	Typicality scale anchors: Medium	Typicality scale anchors: Low	Opt-outs
Goodness-of-example	low to high	in scale	Rosch 1975 (p. 198) Uyeda, Mandler 1980 (p. 588) Brown, Semrau 1986 (p. 124)	1 very good example of your idea of what the category is	4 fits your idea moderately well	7 fits very poorly (or is no member at all)	
		opt-out	Schröder, Gemballa 2012 (p. 383)	1 very good example of the category / typical		7 bad example of the category / atypical	·not a category member ·unfamiliar
	high to low	in scale	Malt, Smith 1984 (p. 260) Schwanenflugel, Rey 1986 (p. 152)	7 very good example of your idea of what the category is	4 fits your idea moderately well	1 fits very poorly (or is no member at all)	
			Moreno-Martinez et al. 2014 (p. 1095)	5 this is a very good example of the category		1 this is a bad example of the category	0: not a category member or not familiar with the item
		not in scale	Barsalou 1985 (p. 634)	9 excellent example 10 the instance is about as good an example as you can get of your idea or image of what the category is		1 poor example 0 you think the item does not fit at all with your idea or image of the category	
			Smith et al. 1988 (p. 502)				
			De Deyne et al. 2008 goodness (p. 1034)	20 very good example		1 very bad example	unknown
Typicality	low to high	in scale	Hampton, Gardiner 1983 (p. 494)	1 very typical instance of the category		5 very atypical instance borderline cases which sometimes belong to a category, but not always	· 6: item does not be-long to the category · unknown
	high to low	not in scale	McCloskey, Glucksberg 1978 (p. 464)	10 the candidate exemplar is highly typical of the category		1 the exemplar is extremely atypical (i.e. unrelated)	unknown
			Ruts et al. 2004 (p. 510)	20 very typical item		1 very atypical or unrelated item	unknown
			De Deyne et al. 2008 typicality (p. 1032)	20 very typical item		1 very atypical item	unknown

3.2.3 **Instructions**

The most prominent instructions are from the first typicality study for semantic categories, Rosch (1975b):

> This study has to do with what we have in mind when we use words which refer to categories. Let's take the word red as an example. Close your eyes and imagine a true red. Now imagine an orangish red ... imagine a purple red. Although you might still name the orange red or the purple red with the term red, they are not as good examples of red (as clear cases of what red refers to) as the clear "true" red. In short, some reds are redder than others. The same is true for other kinds of categories. Think of dogs. You all have some notion of what a 'real dog,' a 'doggy dog' is. To me a retriever or a German shepherd is a very doggy dog while a Pekinese is a less doggy dog. Notice that this kind of judgment has nothing to do with how well you like the thing; you can like a purple red better than a true red but still recognize that the color you like is not a true red. You may prefer to own a Pekinese without thinking that it is the breed that best represents what people mean by dogginess.
>
> On this form you are asked to judge how good an example of a category various instances of the category are. At the top of the page is the name of a category. Under it are the names of some members of the category. After each member is a blank. You are to rate how good an example of the category each member is on a 7-point scale. A 1 means that you feel the member is a very good example of your idea of what the category is. A 7 means you feel the member fits very poorly with your idea or image of the category (or is not a member at all). A 4 means you feel the member fits moderately well. For example, one of the members of the category fruit is apple. If apple fit well your idea or image of fruit, you would put a 1 after it; if apple fit your idea of fruit very poorly you would put a 7 after it; a 4 would indicate moderate fit. Use the other numbers of the 7-point scale to indicate intermediate judgments. Don't worry about why you feel that some-thing is or isn't a good example of the category. And don't worry about whether it's just you or people in general who feel that way. Just mark it the way you see it." (Rosch, 1975b, p. 198)

Her instructions explicitly distinguish liking something and its goodness-of-example (GOE). A seeming contradiction is that typicality is explained as "that best represents what people mean" and the later instruction to rate as it is personally seen, without worrying about intersubjective stability. As shown in Table 10, there are two variations in the wording of the questionnaire: most studies follow Rosch and ask for the GOE of SCs, but some directly ask for typicality. Barsalou (1985) asked for GOE because he considered asking directly for typicality to possibly "bias participants towards using frequency of instantiation. 'How good an example' seemed more open ended and less demanding." (p. 634). Only five of the fourteen investigated studies use the word typicality in the instructions. Ruts et al. (2004) used "standard instructions for typicality" but do not specify which ones. The same goes for McCloskey and Glucksberg (1978). Hampton and Gardiner (1983) ask to rate typicality and state that they followed the instructions used in former research but modified them

> "[…] to make the task clearer and less ambiguous for participants: (a) participants were given a separate rating response for denying that an item belonged in the category, (b) they could also leave a line blank if they did not know a word, and (c) instructions stressed that frequency of occurrence should not be used as a basis for the judgement" (Hampton & Gardiner, 1983, p. 493).

They used the following example to show that frequency does not matter for typicality:

> "For instance, most people would say that Churches are very typical examples of the category Buildings; more typical than, say, Telephone boxes, which some people would classify as very atypical examples. The above example also serves to illustrate the fact that, just because a specific word is more typical than another, it does not mean that it occurs more often in your experience than an atypical word. Telephone boxes are probably seen much more often than Churches. but they are still less typical of the category Buildings than Churches are." (Hampton & Gardiner, 1983, pp. 493–494)

Kittur et al. (2006) found that for artificial categories, typicality and GOE questions lead to different category structures:

> "[…] when categories involve relations, two distinct measures become available: how well an exemplar fits a relational ideal, and the closeness of its features to the central tendency of the category distribution. In this case it is possible that typicality and GOE judgments are not identical, and instead measure different types of graded structure." (Kittur et al., 2006, p. 430)

De Deyne et al. (2008) collected data for both phrasings of the question and found correlations between .82 and .99 (mean = 0.93, SD = 0.06). For fruit, the correlation is .95 and for vegetable .97, which indicates a negligible difference. Hampton (2017) argues that "asking about 'goodness' leads to an evaluative judgement and hence allows ideals to have a greater influence on the judgements" (Hampton, 2017, p. 106). To assess the difference between the two types of wording, I compared the GOE and typicality data from the De Deyne et al. (2008) datasets[5].

Table 11 compares the ratings for fruit and vegetables. Typicality ratings tend to be higher and have a lower variance than GOE ratings, the mean difference of means being .8 for fruit and 1.8 for vegetable and the mean SD difference being -0.7 for fruit and -1.2 for vegetable. The difference of means is particularly high for *pumpkin* (4.6) and *fig* (4.1), which received medium typicality ratings around 12, but GOE ratings in the lower half of the scale with 5.7 and 4.9. However, the SD is very high for both SCs (4.9-6.2), indicating low intersubjective agreement and that the mean might not reflect the data well.

[5] Available at https://ppw.kuleuven.be/apps/concat/datasets/brm_concepts/.

Table 11: Mean typicality (TYP) and goodness (GDN) ratings with SDs and their differences for a) fruit and b) vegetables from De Deyne et al. (2008).

a)

subcategory	Mean TYP	SD TYP	Mean GDN	SD GDN	Mean TYP - GDN	SD TYP-GDN
banana	19.1	1.1	19.0	2.0	0.2	-0.9
apple	19.0	1.4	19.4	1.3	-0.4	0.1
orange	18.9	1.4	18.9	1.8	0.0	-0.5
strawberry	18.3	2.0	18.1	2.1	0.1	-0.1
pear	18.2	2.3	18.6	2.0	-0.4	0.3
grape	17.5	2.6	17.3	3.5	0.2	-0.9
pineapple	17.4	2.8	16.4	3.4	1.0	-0.6
cherry	17.2	2.5	17.1	2.9	0.1	-0.4
kiwi	17.1	2.9	16.7	4.1	0.4	-1.2
peach	17.0	2.8	16.6	3.7	0.4	-0.9
plum	16.7	3.2	15.9	4.6	0.7	-1.4
melon	16.5	3.2	15.7	4.8	0.8	-1.6
raspberry	16.3	3.3	14.8	4.3	1.5	-1.1
nectarine	16.3	3.4	16.6	3.2	-0.4	0.3
apricot	16.0	3.1	15.8	3.9	0.2	-0.8
mandarine	15.8	3.5	16.1	4.7	-0.2	-1.2
grapefruit	15.7	3.7	14.9	4.5	0.8	-0.8
lemon	15.3	3.8	14.1	6.8	1.1	-2.9
mango	15.0	3.6	14.9	4.7	0.1	-1.1
blueberry	15.0	3.9	13.1	6.0	1.9	-2.1
passion fruit	14.9	4.4	14.7	5.3	0.2	-0.9
blackberry	14.4	4.7	14.4	4.8	0.0	-0.1
lime	13.9	4.9	13.6	5.5	0.4	-0.7
lychee	13.5	4.1	12.5	5.8	0.9	-1.8
pumpkin	12.5	6.2	7.9	5.7	4.6	0.6
papaya	12.5	5.0	12.1	5.3	0.4	-0.2
fig	12.4	4.0	8.3	4.9	4.1	-0.8
coconut	12.4	5.2	10.8	4.7	1.6	0.5
red currant	11.8	5.1	11.0	5.7	0.8	-0.5
dates	10.9	4.8	8.4	4.9	2.4	-0.1
mean	15.6	3.5	14.8	4.2	0.8	-0.7

b)

subcategory	TYP	SD TYP	GDN	SD GDN	Mean TYP - GDN	SD TYP-GDN
lettuce	18.9	1.5	17.4	4.8	1.5	-3.3
carrot	18.9	1.3	17.4	4.0	1.4	-2.7
leek	18.3	2.4	15.8	5.5	2.4	-3.1
cauliflower	18.0	2.5	16.6	3.8	1.4	-1.3
spinach	18.0	2.2	16.3	3.8	1.8	-1.7
beans	18.0	2.0	15.2	4.7	2.8	-2.7
endive	17.7	3.1	15.7	4.9	2.0	-1.8
tomato	17.3	5.0	16.9	4.6	0.4	0.4
peas	17.2	3.2	15.4	5.0	1.8	-1.8
brussels sprouts	17.1	3.9	15.7	4.6	1.4	-0.7
red cabbage	17.1	2.6	14.9	5.1	2.2	-2.6
white cabbage	17.0	2.9	15.0	4.8	1.9	-1.9
cucumber	16.8	3.4	16.2	4.5	0.6	-1.2
celery	16.7	3.9	14.7	5.8	2.0	-1.9
asparagus	16.4	4.4	14.4	4.9	2.0	-0.4
zucchini	15.8	4.6	13.9	5.1	1.9	-0.5
eggplant	15.7	4.8	12.9	5.7	2.9	-0.9
pepper	15.1	4.4	13.8	6.0	1.4	-1.6
black salsify	14.7	4.3	13.0	5.5	1.7	-1.2
radish	14.5	4.4	12.4	5.6	2.2	-1.3
beet	14.0	4.0	10.4	5.7	3.7	-1.7
mushrooms	13.5	5.5	11.2	6.1	2.4	-0.6
chervil	13.5	5.1	11.0	6.1	2.5	-1.0
corn	13.0	4.6	11.2	5.0	1.8	-0.4
onions	12.9	4.8	11.1	6.0	1.7	-1.2
gherkins	12.8	5.3	10.9	4.9	2.0	0.4
water cress	12.5	5.0	10.7	5.4	1.9	-0.3
parsley	12.0	6.2	10.7	5.7	1.3	0.4
potato	10.7	6.3	10.0	6.5	0.7	-0.2
garlic	9.5	5.5	7.8	5.2	1.7	0.2
mean	15.5	4.0	13.6	5.2	1.8	-1.2

Figure 15 compares the mean typicality and goodness ratings, their SDs as well as the frequency distributions for all 16 categories used in De Deyne et al. (2008). Both relationships are linear, but mean goodness has a trend to be lower than mean typicality on the lower end of the scale, which can be seen in Figure 15a), and the goodness ratings tend to have a higher SD, which can be seen in Figure 15 b). Figure 15c) and d) show the rating distributions for typicality resp. GOE ratings. In both question types, 20 is the most frequent and 2 the least frequent response. Ratings under 10 are used more frequently when asking for GOE and the lowest rating 1 is used almost twice as often than when asking for typicality.

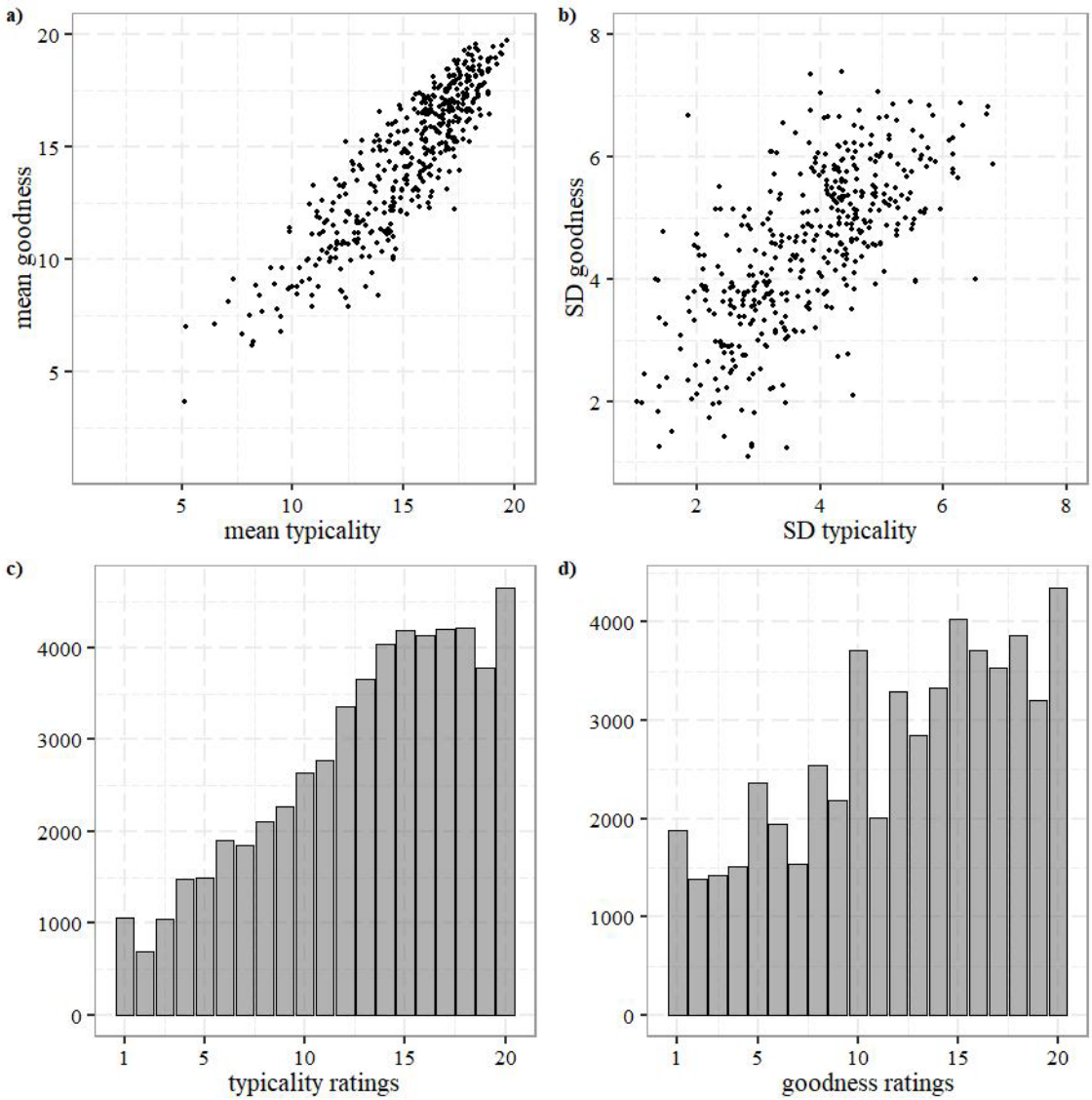

Figure 15: Comparison between typicality and goodness for all categories reported in De Deyne et al. (2008). a) mean typicality against mean goodness ratings, b) SDs of mean typicality ratings and mean goodness ratings, c) frequency distribution of typicality ratings, d) frequency distribution of goodness ratings.

This comparison shows that while the wording seems interchangeable, as both datasets are almost perfectly correlated, it is not necessarily when the actual mean values are important. GOE instructions seem to lead to a more frequent use of the lower scale points and higher SDs.

3.2.4 **Participants**

The information on the participants taking part in the typicality studies is summarised in Table 12. Most studies chose students as participants, often undergraduate and psychology students. Age and gender are not consistently reported. Most data are from English speakers, but two Dutch, two Spanish and one German study are available. Furthermore, 7 of the 9 English studies were carried out in the USA, additionally one from England and one from Ireland are available. Moreno-Martínez et al. (2014) had the additional criterion of normal or corrected-to-normal vision, and Spanish to be the participants' first language and exclusion of people who had suffered neurological traumata (p. 1090). Schwanenflugel and Rey (1986) had the additional criterion of living in Florida with a mean length of 14.9 years for the English speakers and 16.4 years for the Spanish speakers. The number of participants who made ratings for each category

varies between 10 to 209, with 9 of the 14 studies employing between 20 and 51 participants, 2 fewer than 20 and 2 considerably more, 120 to 160 and 209. It is an open question how many ratings are required for reliable results. At least 30 is a rule-of-thumb. Brown and Semrau (1986) point out that using few participants makes the data less reliable and state that studies with few participants should not be regarded as norms:

> "McCloskey and Glucksberg (1978) […] employed less than 25 raters which may make their data less reliable than ours. Malt & Smith (1982) and Barsalou (1985) used no more than 20 raters; Caramazza, Hersh & Torgerson (1976) used 38. But all these authors might wish their data to be regarded as published experimental findings rather than as norms with the implications of accuracy and stability which such a term carries." (p. 128-130)

In the article "The weirdest people in the world" (Henrich et al., 2010), doubt is shed on the generalizability of samples drawn exclusively from Western, Educated, Industrialized, Rich, Democratic (WEIRD) population. Rosch (1975b) notices this point and argues that, for her claim to be valid, the intersubjective stability of ratings does not have to hold between all cultures – "no claim is made that the internal structure of semantic categories should be universal for all cultural groups." (p. 199). She stresses the necessity of taking the same population for typicality and derived measures: "the population on which these norms were collected is the same population on which further experiments using the norms as an independent variable were performed." (p. 199).

Table 12: Summary of participant information in typicality studies (NA means no information available).

dataset	age	education status	location	gender	language	participants per category
Rosch 1975	NA	college students enrolled in psychology classes	Berkely, California, USA	NA	English natives	209
McCloskey and Glucksberg 1978	NA	undergraduate students	Princeton, New Jersey (Princeton University), USA	NA	English	24
Uyeda and Mandler 1980	NA	undergraduate students enrolled in introductory psychology courses	San Diego, California, USA	50 male 50 female	English natives	50
Hampton and Gardiner 1983	NA	students, about half psychology students	London, England	NA	English	43-51
Malt and Smith 1984	NA	students	Stanford, California, USA	NA	English	16-19
Barsalou 1985	NA	students	Stanford, California, USA	NA	English	10
Brown and Semrau 1986	NA	first-year psychology students	Belfast (Queen's University), Ireland	half male	English	120-160
Schwanenflugel and Rey 1986	NA	NA	South Florida, USA	27 female 23 male	English Spanish	50
Smith et al. 1988	NA	undergraduate students	Cambridge (Harvard University), Massachusetts, USA	NA	English	30
Ruts et al. 2004	NA	research assistants, last-year psychology students (89) and first-year psychology students (225)	Leuven (University of Leuven), Netherlands	NA	Dutch	21-25
De Deyne et al. 2008	18-63 (M=20,5)	second-year psychology students	Leuven (University of Leuven), Netherlands	89 female 23 male	Dutch	28
Schröder and Gemballa 2012	23-69 (M=45.05, SD=17.42)	some students, years of education: 10-13 (M=12.05, SD=1.05)	Germany	15 female 5 male	German monolingual and native	20
Moreno-Martinez 2014	19-65 (M=33.7, SD=10.5)	undergraduate students	Spain	76 male 76 female	Spanish	38
Hampton French data	NA	NA	France			55

3.2.5 Comparison of all datasets

First, the datasets had to be combined and made comparable. This was done by

- transforming all SCs into the singular form (*baked beans*, *chives*, and *cloves* were left in plural, because the singular does not seem to be in common use),
- using the American English variant where they differ (*aubergine* to *eggplant*, *courgette* to *zucchini*, *gherkin* to *pickle*),
- merging synonyms and different spellings (*beetroot* to *beet, water cress* to *watercress, honeydew melon* to *honeydew, musk melon* to *muskmelon, water melon* to *watermelon, mandarine* to *mandarin, litchi* to *lychee, scallion* and *spring onion* to *green onion),*
- correction of obvious typos (lima to *lime,* fench beans to *French beans*),
- translation of the French subcategories (kindly done by a native speaker).

In total, mean ratings are available for 104 different SCs for the fruit category and 120 for the vegetable category.

Moreno-Martínez et al. (2014) report data for 5 different kinds of peach[6]. After correspondence with the author, two of them (*abridero* and *albérchigo*) had to be excluded due to a transcription error. Then, after some research and comparison with the other Spanish dataset from Schwanenflugel and Rey (1986), *melocotón* was translated as *peach*, *paraguaya* as *Saturn peach* and *fresquilla* was kept in the Spanish original. They also report 3 different kinds of cherry.[7] As typicality was very high for cherries in the other datasets, the one with the highest mean typicality rating, *cereza*, was used as cherry and the other two were kept in the Spanish original.

Then, the reported mean ratings of SCs were normalised using the formula

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}},$$

Resulting in values between 0 and 1, where 1 is the highest possible typicality rating. The reported rating of 1.07 for *orange* from Rosch (1975b) on a 7-point-scale where 7 corresponds to minimal typicality is for example normalised as $\frac{1.07-7}{1-7} = 0.99$.

In addition, the meta mean (the mean of the means) and meta SD (the SD of the means) were calculated for all SCs that were used in more than one study. For the meta SD, the whole-group SD was calculated, because the datasets available are the whole population of available values, thus the sum of the squared differences was divided by N, not N-1. Figure 16 shows the meta means and SDs for all SCs used in at least 3 datasets. It can be seen that the SDs between datasets are lower on the high end of the typicality scale and higher on the medium and low typicality level.

[6] abridero (translated as "type of peach", familiarity of 0.55 and typicality of 0.47), albérchigo (translated as "Clingstone peach", familiarity of 0.69 and typicality of 0.5), fresquilla (familiarity of 2.15 and typicality of 1.83) and melocotón (familiarity of 4.55 and typicality of 4.55) which are both translated as "Peach", and paraguaya (translated as "Saturn peach").

[7] Cereza (typicality 4.32 and familiarity 4.11), guinda (typicality 2.94, familiarity 3.04) and picota (typicality 2.9, familiarity 3.02).

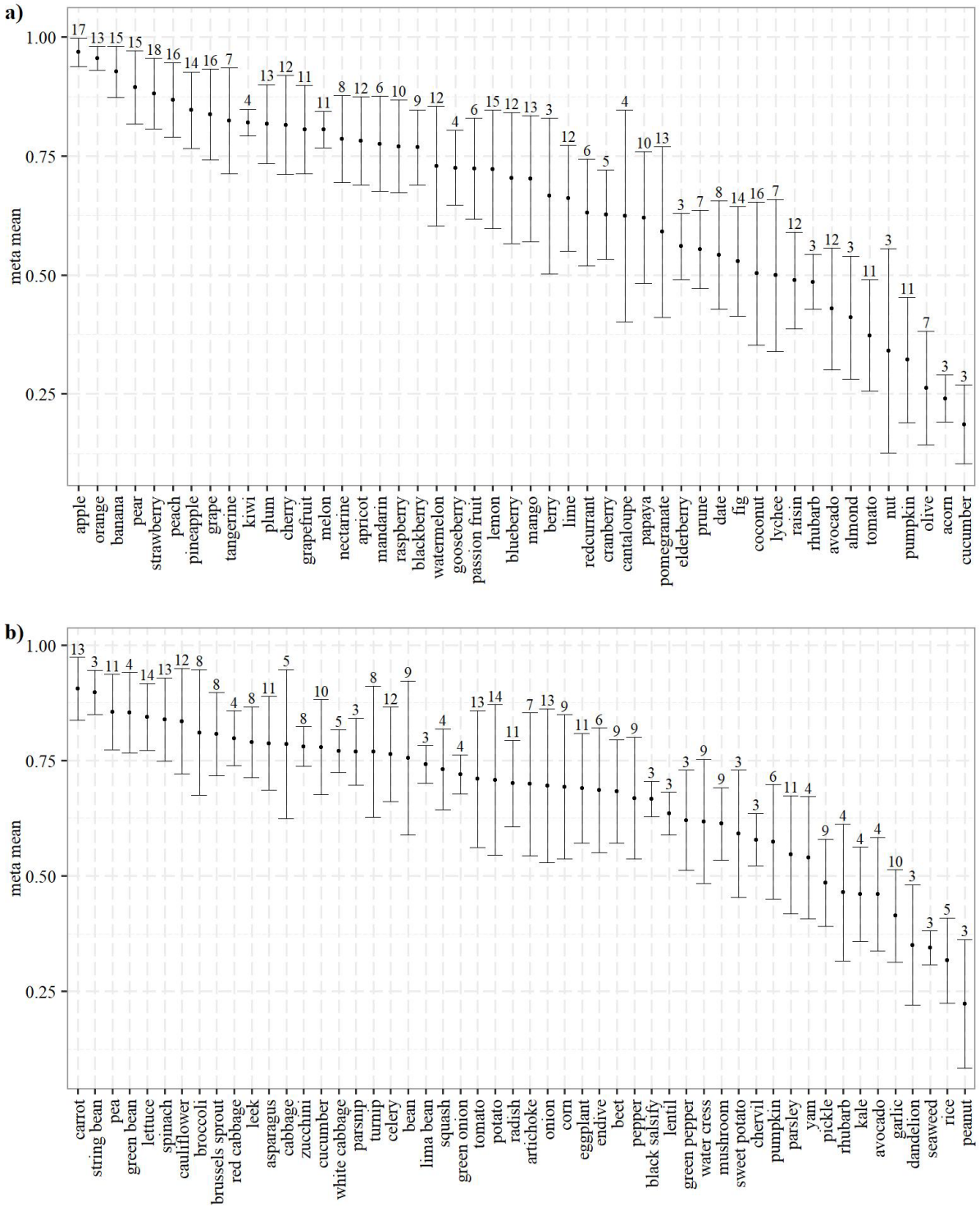

Figure 16: Meta means with error bars reflecting the SD and number of studies for SCs used in 3 or more studies for a) fruit and b) vegetables. The number of studies is on top of the whiskers

Figure 17 shows the distribution of normalised typicality means for each study in histograms. A uniform distribution would include similar amounts of SCs for all typicality levels. Then, all bars would have similar heights and the skew would be close to 0. All datasets have a negative skew with more SCs on the high typicality level, except the French dataset for both categories and the datasets from McCloskey and Glucksberg (1978) and Moreno-Martínez et al. (2014)

for *fruit*. Many do not include SCs on the lowest scale intervals and those that do only contain few.

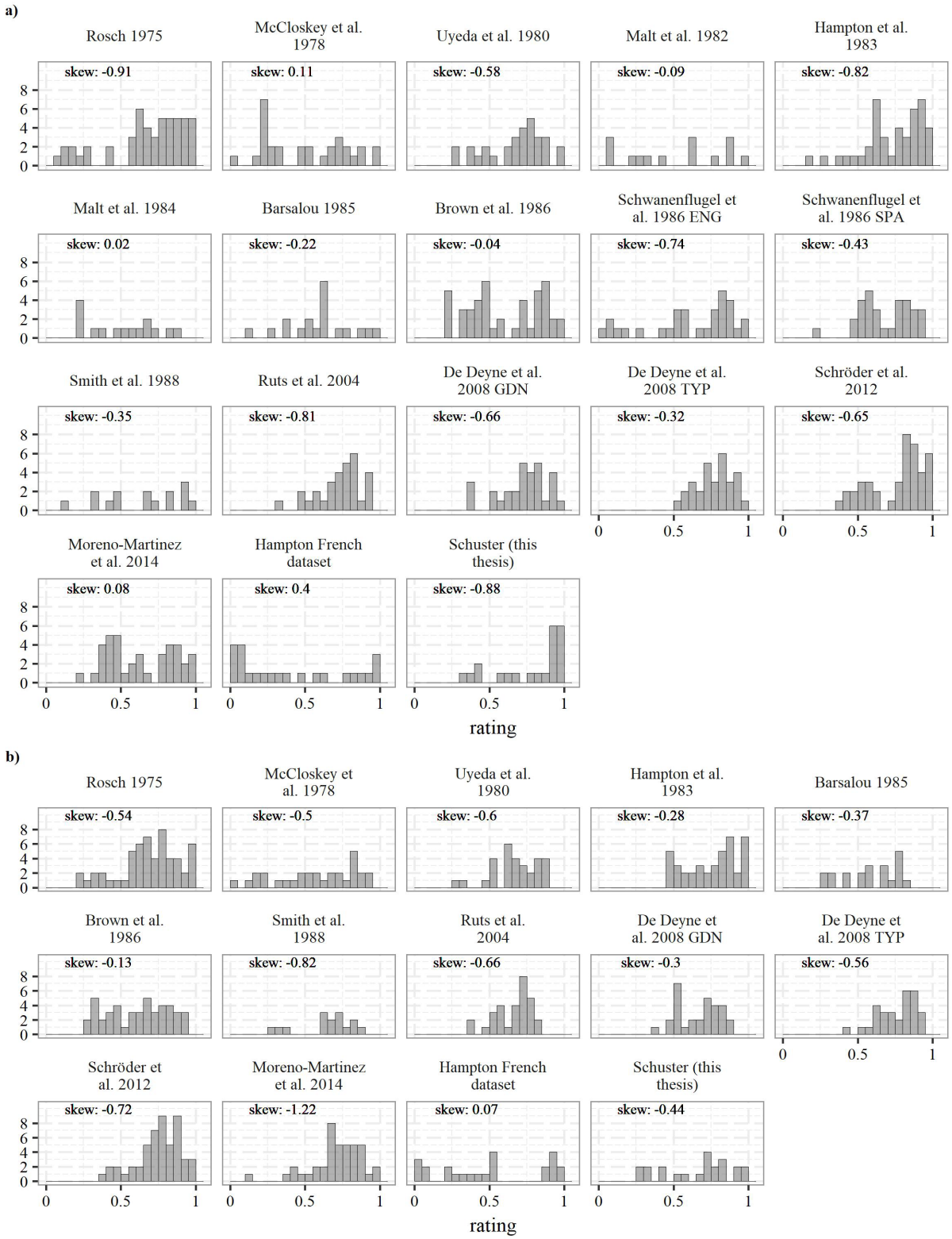

Figure 17: Histograms of normalised mean typicality ratings for SCs per study for a) fruit and b) vegetables.

Figure 18 shows for fruit a) the number of studies in which SCs were used, b) the meta mean against the number of studies in which the SCs were used, c) how the meta SD is related to the number of studies and d) the meta SDs against the meta means. Figure 18a) shows that more

than half of the SCs (54%, 56 out of 104) are uniquely used in one or two studies and only 38 SCs (37%) were used in 5 or more studies. From Figure 18b) it becomes clear that the lower end of the typicality scale contains particularly many unique SCs. There is no SC with a meta mean <.25 that is used in more than 8 studies. Regarding the SD between studies, Figure 18c) shows that most SCs' meta means have an SD <.2 and that the highest meta SDs are found for SCs used in a lower number of studies. Figure 18d) shows that the highest meta SDs are at the medium to low meta means. There is one SC with a particularly high meta SD of .39: *mammee* from the English and Spanish Schwanenflugel et al. dataset. While the English speakers rated it as very untypical (normalised mean of 0.04, SD = .15), the Spanish speakers rated it as typical (normalised mean of .81, SD = .3). There are 3 other SCs that were rated very differently by Spanish and English speakers and that consequently have a high meta SD: *soursop* with meta SD of .27 (normalised meta mean .08 vs. .61), *tamarind* with a meta SD of .26 (normalised meta mean .12 vs. .63) and *cherimoya* with a meta SD of .25 (normalised meta mean .07 vs. .58). *Nut* and *cantaloupe* have SDs greater than .2, the SDs for the remaining SCs are below .2.

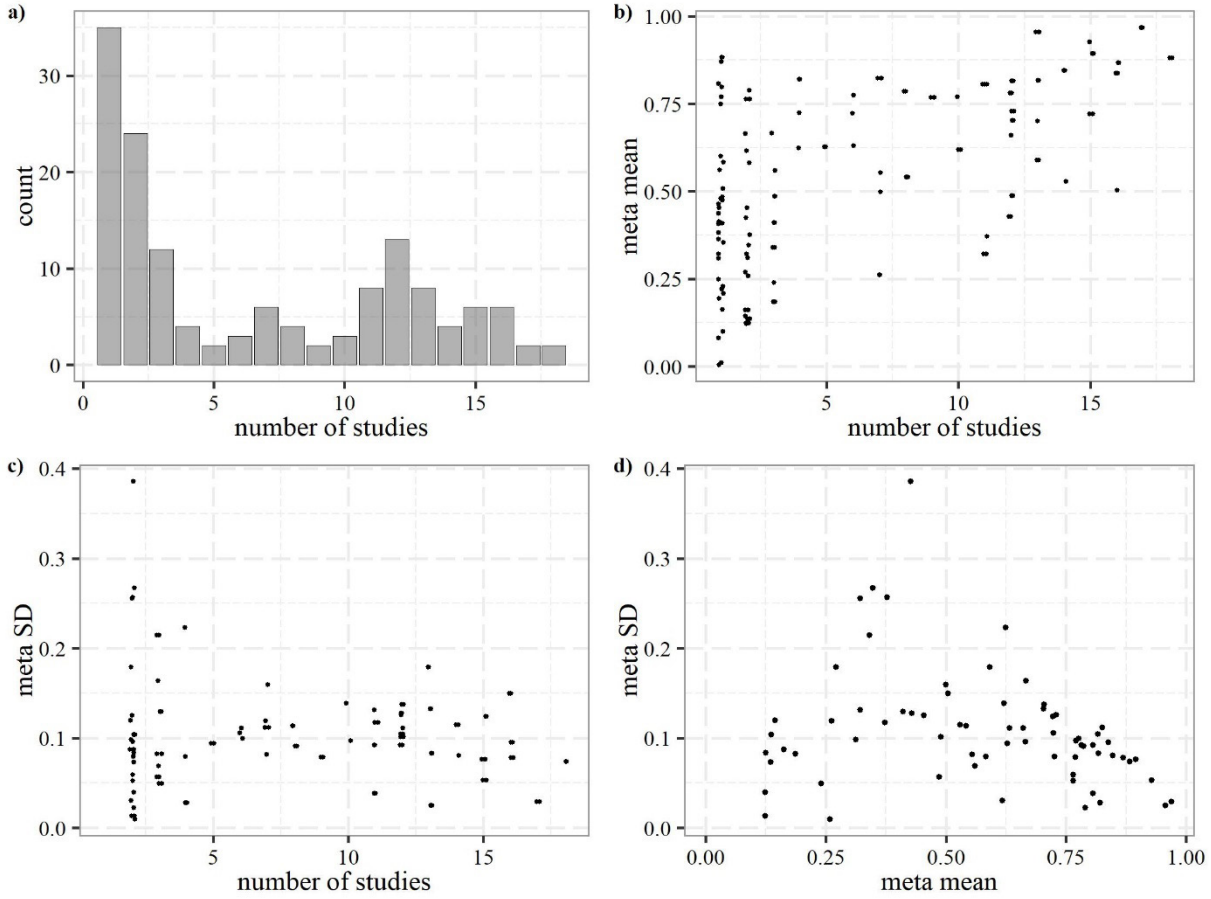

Figure 18: Fruit: a) Histogram of number of studies in which SCs are used, b) number of studies against meta mean (jitter width = 0.1), c) number of studies against meta SD for all SCs used in 3 or more studies (jitter width = 0.1), and d) meta mean against meta SD.

Figure 19 shows the same data for vegetable. Like for fruit, it can be seen in Figure 19a) that more than half of the SCs are used uniquely in one or two studies (68 SCs, 57%) and only 33 SCs (28%) were used in 5 or more studies. Figure 19b) shows that of those SCs used in many studies, most have a meta mean of .5 or higher and there is no SC with a meta mean < .25 used in more than 3 studies. Different from the fruit category, most meta SDs are below .2 and there is no systematic relationship between meta SD and the number of studies (Figure 19c)).

The higher meta SDs are more evenly distributed among the meta means (Figure 19d)). There is one SC with a very high SD of .34: *sprouts* was rated to be very typical in Hampton and Gardiner (1983) (normalised mean .97) and untypical in Barsalou (1985) (normalised mean of .3). All other meta SDs are ≤.17.

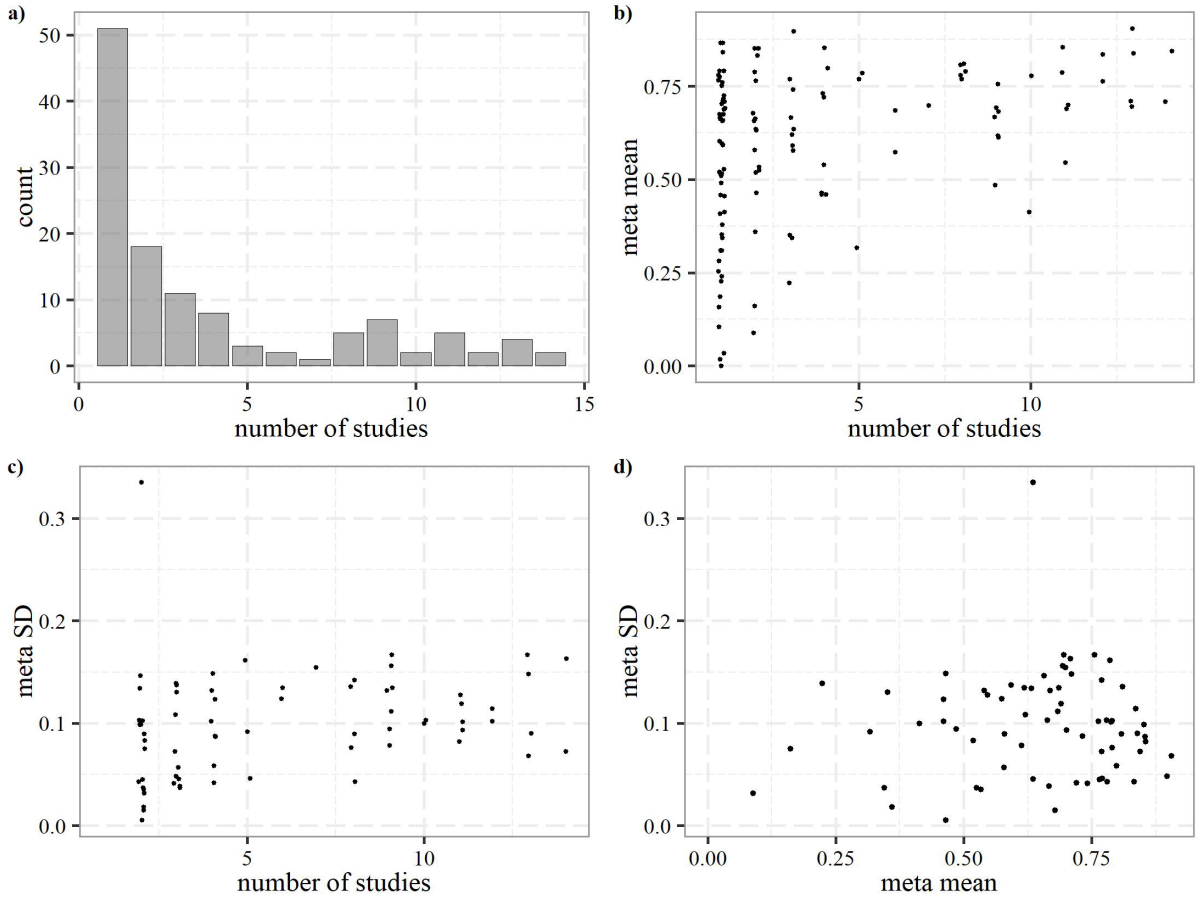

Figure 19: Vegetable: a) Histogram of number of studies in which SCs are used, b) number of studies against meta mean (jitter width = 0.1), c) number of studies against meta SD for all SCs used in 3 or more studies (jitter width = 0.1), and d) meta mean against meta SD for all SCs used in 3 or more studies.

Figure 20 shows the distribution of meta-means per typicality level for all studies. In both categories, there are almost no SCs with a meta mean ≥ .9, for fruit there are three (*orange, apple* and *banana*) and for vegetable one (*carrot*). There are also few SCs with a meta mean < .1, three for *fruit* (*chicken, ginger* and *mushroom*) and four for *vegetable* (*bread, steak, milk, pineapple*). This part of the scale seems to be reserved for clear-non-members, which most studies did not include. The other intervals are relatively evenly covered for fruit: each contains between 8% and 18% (8 to 18) SCs. The *vegetable* data has a negative skew: 57 SCs (48%) have a meta mean between .6 and .8 and there are only 4-5 SCs between .1 and .2 and between .2 and .3. In the remaining intervals there are around 10 SCs each.

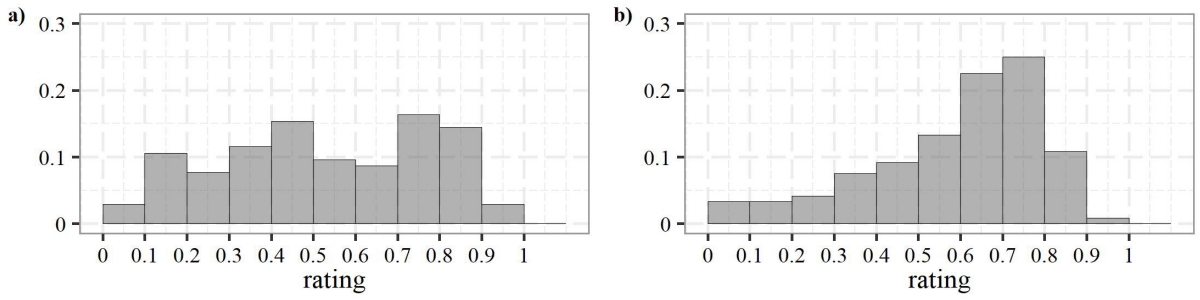

Figure 20: Distribution of meta means for a) fruit and b) vegetables.

3.2.6 **Intersubjective stability**

It is commonly assumed that typicality is an intersubjectively stable psychological variable. This means that the correlations of the mean ratings between the different studies should be high. The correlations might be influenced by the fact that the criteria of SC selection differ between studies (section 3.2.1) and there are not many SCs that were used in many studies, in particular on low typicality levels (section 3.2.5).

The correlation matrix for *fruit* is in Table 13. The mean correlations of the datasets are between .72 and .91 for fruit (mean = .85, SD = .05). The original experiment reported in Rosch (1975b) was used as a basis of many of the further studies. The correlations of her experiment with other studies are very high, between .81 and .98 (all $p < .001$), except for the ones with the two Spanish datasets, which are .66 and .67 ($p < .001$). It is also remarkable that the degree of overlap (i.e., common SCs) with the other studies is high (between 12 and 31). For the other studies, the number of pairs for comparison is partly very low which is due to the high variability in the SC selection between studies. As shown in section 3.2.5, this mainly affects SCs on the medium and low typicality level. It was also seen that this is the level with most intersubjective variance in the ratings, while participants agree in general about good examples. Therefore, the high correlations should be interpreted with caution about confirming the general stability of typicality orderings. What they do confirm is that participants show a high agreement about which SCs are typical category members.

The correlation matrix for *vegetable* is in Table 14. They are between .31 and .78 (mean = .6, SD = .13). The data from Rosch (1975b) again has high correlations between .86 and .97 ($p < .001$) with data from the USA and with the French dataset. The correlations range between .42 and .71 with the datasets from England, Ireland, the Netherlands, Germany and Spain. High correlations for the same language or close cultural spaces are found in all cases. The two studies from Great Britain, Hampton and Gardiner (1983) and Brown and Semrau (1986), have a high correlation of .85 ($p < .001$) and the three Dutch datasets have correlations between .91 and .97, $p < .001$. The single studies available in German and Spanish have low correlations with all other studies with a mean of .31 resp. .41. This indicates an interculturally different perception of vegetables, which is presumably related to different diets.

Table 13: Correlations matrix for fruit with significance levels and number of common SCs in brackets.

	McCloskey et al. 1978	Uyeda et al. 1980	Hampton et al. 1983	Malt et al. 1982	Malt et al. 1984	Barsalou 1985	Brown et al. 1986	Schwanenflugel et al. 1986 ENG	Schwanenflugel et al. 1986 SPA	Smith et al. 1988	Ruts et al. 2004	De Deyne et al. 2008 TYP	De Deyne et al. 2008 GDN	Schröder et al. 2012	Moreno-Martinez et al. 2014	Hampton French dataset	Schuster (this thesis)	mean
Rosch 1975	0.98*** (18)	0.93*** (30)	0.85*** (34)	0.94*** (15)	0.91*** (15)	0.81*** (19)	0.82*** (33)	0.93*** (24)	0.66*** (24)	0.96*** (15)	0.84*** (26)	0.8*** (26)	0.84*** (26)	0.8*** (31)	0.67*** (26)	0.82*** (12)	0.91*** (21)	0.86
McCloskey et al. 1978		0.96*** (13)	0.92*** (15)	0.93*** (8)	0.91*** (10)	0.94** (7)	0.72** (16)	0.96*** (11)	0.87*** (11)	0.92*** (9)	0.94*** (9)	0.89** (9)	0.93*** (9)	0.89*** (10)	0.73** (14)	0.94*** (15)	0.9*** (13)	0.91
Uyeda et al. 1980			0.87*** (27)	0.99*** (11)	0.97*** (13)	0.87*** (19)	0.83*** (26)	0.99*** (23)	0.75*** (23)	0.94*** (14)	0.88*** (21)	0.92*** (21)	0.9*** (21)	0.9*** (23)	0.82*** (24)	0.71 (8)	0.84*** (17)	0.89
Hampton et al. 1983				0.94*** (13)	0.89*** (15)	0.73*** (17)	0.96*** (36)	0.88*** (23)	0.65*** (23)	0.81*** (13)	0.83*** (25)	0.81*** (26)	0.84*** (26)	0.83*** (29)	0.78*** (25)	0.74** (12)	0.8*** (18)	0.84
Malt et al. 1982					0.97*** (13)	0.92*** (9)	0.95*** (11)	0.99*** (8)	0.81* (8)	0.98*** (11)	0.95*** (10)	0.98*** (10)	0.96*** (10)	0.95*** (9)	0.88*** (10)	0.73 (5)	0.93*** (9)	0.93
Malt et al. 1984						0.77* (9)	0.92*** (13)	0.97*** (10)	0.81** (10)	0.85*** (13)	0.94*** (10)	0.91*** (10)	0.92*** (10)	0.94*** (10)	0.88*** (12)	0.69 (7)	0.79** (11)	0.89
Barsalou 1985							0.85*** (15)	0.92*** (14)	0.85*** (14)	0.88*** (10)	0.81*** (14)	0.87*** (14)	0.87*** (14)	0.62* (16)	0.82*** (16)	0.97** (5)	0.78** (10)	0.85
Brown et al. 1986								0.87*** (25)	0.66*** (25)	0.89*** (12)	0.85*** (24)	0.83*** (25)	0.85*** (25)	0.84*** (27)	0.79*** (24)	0.7** (13)	0.79*** (18)	0.84
Schwanenflugel et al. 1986 ENG									0.52** (31)	0.95*** (11)	0.78*** (20)	0.79*** (21)	0.76*** (21)	0.83*** (20)	0.81*** (22)	0.72 (7)	0.85*** (15)	0.86
Schwanenflugel et al. 1986 SPA										0.88*** (11)	0.58** (20)	0.6** (21)	0.66** (21)	0.49 (20)	0.67*** (22)	0.75 (7)	0.72** (15)	0.72
Smith et al. 1988											0.91*** (9)	0.94*** (9)	0.91*** (9)	0.81** (10)	0.84*** (13)	0.78 (6)	0.97*** (11)	0.90
Ruts et al. 2004												0.92*** (29)	0.97*** (29)	0.81*** (27)	0.82*** (21)	0.81 (6)	0.87*** (16)	0.86
De Deyne et al. 2008 TYP													0.95*** (30)	0.84*** (26)	0.88*** (21)	0.55 (6)	0.76*** (16)	0.85
De Deyne et al. 2008 GDN														0.8*** (26)	0.79*** (21)	0.71 (6)	0.86*** (16)	0.86
Schröder et al. 2012															0.9*** (22)	0.72 (6)	0.73** (16)	0.82
Moreno-Martinez et al. 2014																0.62 (9)	0.57* (17)	0.79
Hampton French dataset																	0.81** (9)	0.77
Schuster (this thesis)																		0.83

*p<0.02,**p<0.01,***p<0.001

Table 14: Correlations matrix for vegetables with significance levels and number of common SCs in brackets.

	McCloskey et al. 1978	Uyeda et al. 1980	Hampton et al. 1983	Barsalou 1985	Brown et al. 1986	Smith et al. 1988	Ruts et al. 2004	De Deyne et al. 2008 TYP	De Deyne et al. 2008 GDN	Schröder et al. 2012	Moreno-Martinez et al. 2014	Hampton French dataset	Schuster (this thesis)	mean
Rosch 1975	0.92***	0.86***	0.71***	0.89***	0.63***	0.97***	0.59**	0.63***	0.63***	0.54*	0.42	0.9***	0.77***	0.75
	(22)	(29)	(30)	(18)	(34)	(15)	(25)	(25)	(25)	(21)	(29)	(15)	(21)	
McCloskey et al. 1978		0.82**	0.62*	0.57	0.57*	0.85**	0.48	0.48	0.44	0.51	0.1	0.87***	0.73**	0.64
		(12)	(16)	(8)	(18)	(8)	(13)	(13)	(13)	(10)	(13)	(12)	(13)	
Uyeda et al. 1980			0.62**	0.73***	0.48*	0.85***	0.42	0.52	0.5	0.12	0.26	0.78*	0.64*	0.62
			(21)	(17)	(24)	(11)	(19)	(19)	(19)	(18)	(21)	(9)	(13)	
Hampton et al. 1983				-0.05	0.85***	0.86***	0.48*	0.42	0.38	0.62**	0.43	0.88***	0.88***	0.62
				(17)	(31)	(11)	(24)	(24)	(24)	(25)	(27)	(11)	(17)	
Barsalou 1985					-0.17	0.88**	0.28	0.35	0.34	0.3	0.15	0.91*	0.58	0.48
					(16)	(8)	(14)	(14)	(14)	(15)	(16)	(6)	(12)	
Brown et al. 1986						0.81**	0.56**	0.54**	0.58**	0.33	0.37	0.7*	0.59**	0.56
						(11)	(23)	(23)	(23)	(18)	(26)	(11)	(18)	
Smith et al. 1988							0.83**	0.75*	0.76*	-0.12	0.75*	0.96**	0.83**	0.78
							(10)	(10)	(10)	(7)	(9)	(6)	(10)	
Ruts et al. 2004								0.92***	0.91***	0.07	0.41	0.74	0.73**	0.60
								(30)	(30)	(21)	(23)	(9)	(16)	
De Deyne et al. 2008 TYP									0.97***	-0.01	0.49*	0.57	0.61*	0.59
									(30)	(21)	(23)	(9)	(16)	
De Deyne et al. 2008 GDN										0.27	0.56**	0.58	0.58*	0.61
										(21)	(23)	(9)	(16)	
Schröder et al. 2012											0.34	-0.24	0.55	0.31
											(21)	(6)	(14)	
Moreno-Martinez et al. 2014												0.14	0.3	0.41
												(11)	(17)	
Hampton French dataset													0.93	0.69
													(5)	
Schuster (this thesis)														0.69

*p<0.02,**p<0.01,***p<0.001

3.2.7 **Rating behaviour**

Two kinds of data can be analysed to evaluate the participants' rating behaviour in typicality experiments. First, the SDs from those studies which reported them, which were 10 for fruit and 8 for vegetable. Second, the rating distributions from the studies for which the individual ratings are available, Ruts et al. (2004), De Deyne et al. (2008), Hampton's French dataset and my experiment 3.

To compare the SDs between different datasets, the reported values were divided by the maximally possible SD per mean calculated with the Bhatia-Davis-inequality, which states that

$$v \leq (\text{maximum} - \text{mean})(\text{mean} - \text{minimum})$$

(Bhatia & Davis, 2000, p. 353).

This formula describes the fact that different mean values have different maximally possible SDs. For example, if ratings are made on a scale from 1 to 7, a mean of 1 and a mean of 7 have a maximally possible SD of 0, because they can only occur if all ratings are either 1 or 7. From these extreme values the SD rises until it reaches its global maximum value 3 for a mean of 4.

Normalising all SDs with the maximally possible SD corresponding to the means for which they were observed allowed us to compare the SDs of different datasets despite their use of different scales and the resulting different absolute mean values. Figure 21 shows the SDs divided by the maximum SDs against the normalised mean for all studies that reported SDs. Contrary to prior assumptions, these graphs show that there is no difference in intersubjective stability for the different typicality levels. Instead, the SDs are high irrespective of the mean value. The lowest ones are however found for some SCs on the high typicality level.

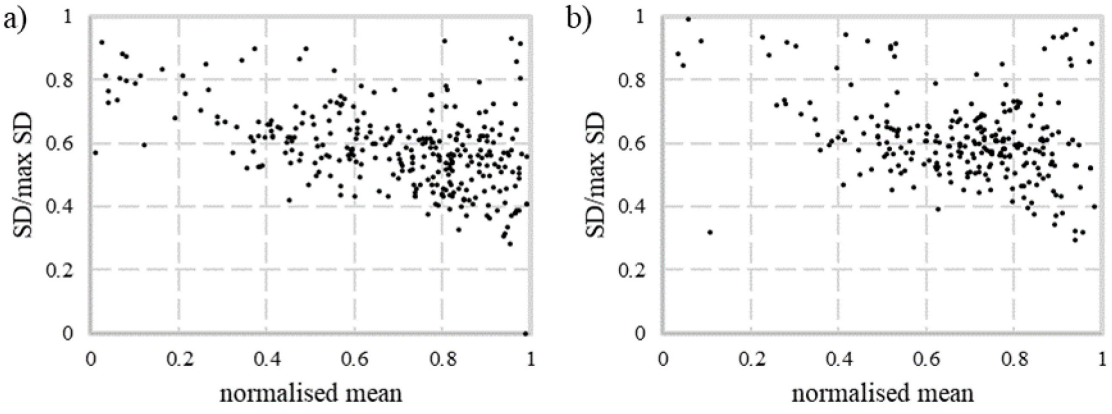

Figure 21: SDs divided by maximum SDs against the normalised mean for all studies that reported SDs for a) fruit and b) vegetable.

In a next step, the four datasets for which individual ratings are available were analysed. Ruts et al. (2004) report ratings for 13 categories by 25 subjects (103 ratings missing) for a total of 344 SCs. De Deyne et al. (2008) made available ratings from 28 subjects for 16 categories (50 ratings missing) for a total of 423 SCs. Hampton kindly provided data for 8 categories with ratings from 55 subjects (45 ratings missing) for a total of 192 SCs. The data from experiment 3 (section 6.3.4) contains ratings by 30 participants for fruit and 32 for vegetables for a total of 44 SCs. In total, there are typicality ratings for 18 different categories, 644 subcategories and 1,002 rating distributions contained in the analysis.

Figure 22 shows the overall rating frequency distribution for all 4 datasets. In all datasets, the highest rating is chosen disproportionally often. In Hampton's French data, the lowest rating occurs almost as often as the highest rating. In Ruts et al. (2004), the ratings 1, 5, 10, and 15 were selected more often than the surrounding ones.

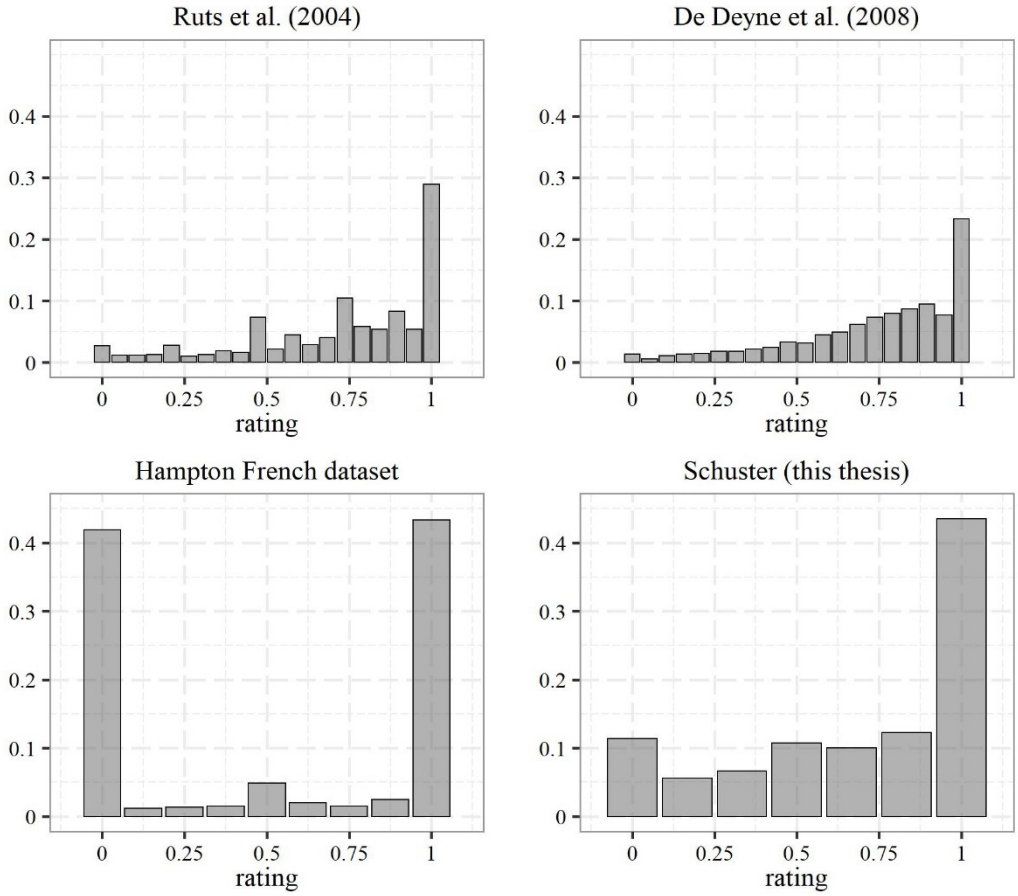

Figure 22: Frequency distribution of ratings for all categories from different studies.

Figure 23 shows the rating distributions for fruit SCs from the De Deyne et al. dataset. To facilitate the discussion of their data, the ratings are discussed in what follows on a reduced scale of 5 scale points, summarising 4 rating points in one (black bars in the figures). The typicality ordering arises not from each SC having a clear mode, but from the ratings being less intersubjectively stable for lower means, like in experiment 3 (see section 6.3.4). Consider *banana* for fruit: 100% of its typicality ratings are between 17 and 20, resulting in the highest observed mean typicality for fruit (19.1). *Dates* has the lowest mean for fruit (10.9), but not because most participants rated the typicality with a value around 11, but because 5 participants each rated it between 17 and 20 and between 13 and 16, 8 rated it between 9 and 12, 7 between 5 and 8 and 3 between 1 and 4. The Ruts et al. data and the French data also show high variances and multimodal distributions for medium typicality levels.

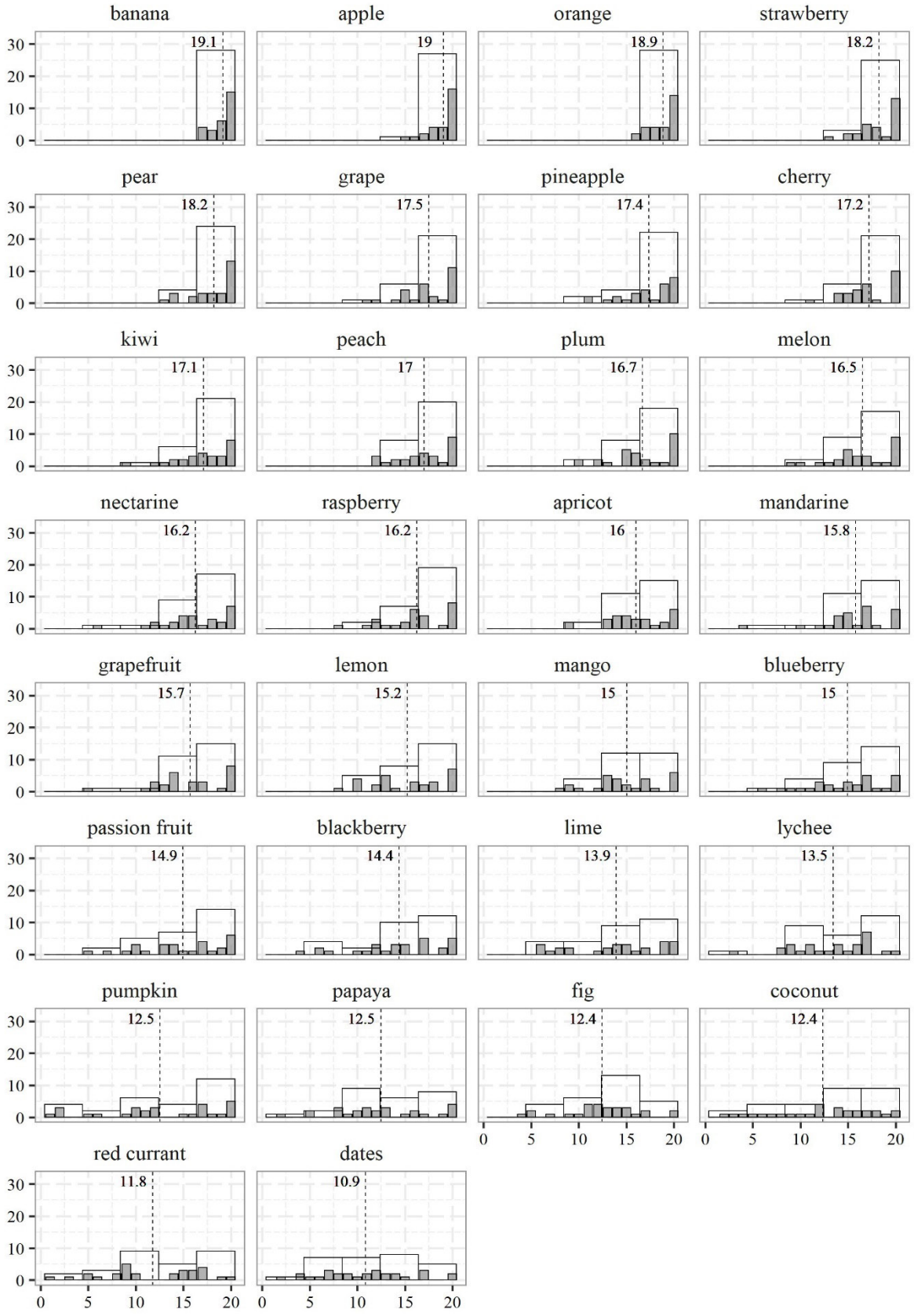

Figure 23: Histogram of typicality ratings for all fruit subcategories from the De Deyne et al. dataset. Grey bars reflect frequency of each individual rating, black-contoured bars reflect the frequency of 4 adjacent scale points combined, the dashed line represents the mean.

The frequent occurrence of multimodal rating distributions led to a further examination of the data in terms of the dip test from Hartigan and Hartigan (1985), as presented in section 3.1.2. A low dip p-value together with a high IQR indicates that there is more than one local maximum within the medium 50% of the data. For comparability, the IQR was normalised between 0 and 1.

Of the 1,002 available rating distributions, 334 (33%) have a dip p-value ≤.05. Figure 24 shows the relationships between a) mean and IQR, b) mean and dip p-value and c) IQR and dip p-value. The IQR is lowest for high and low means and is higher than .5 only between .25 and .75. Low dip p-values which indicate multimodality are found for means between .2 and .98. High IQRs over .8 have low dip p-values exclusively.

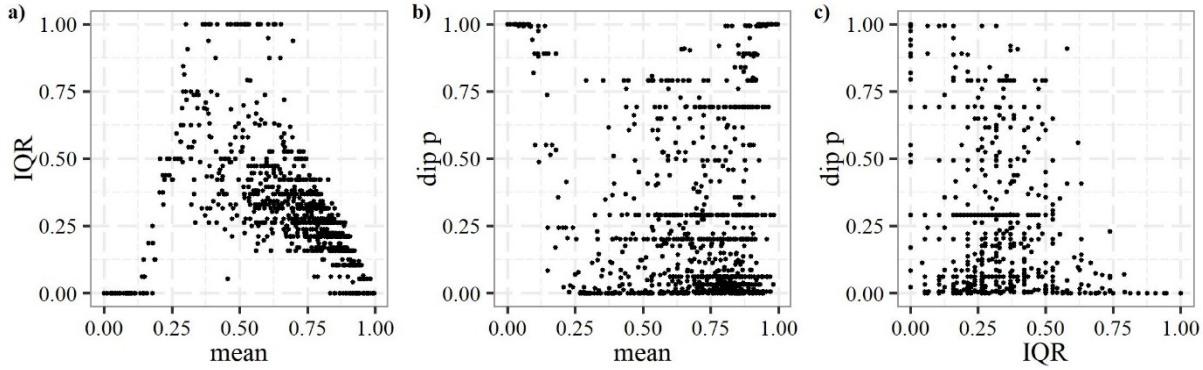

Figure 24: For all datasets and all categories a) normalised mean against normalised IQR, b) normalised mean against normalised p-value from dip-test, c) normalised IQR against normalised p-value from dip-test.

I classified all distributions as multimodal that had a dip p-value ≤ .05 and a high IQR that corresponds to at least 5 scale-points for the 20-point-scales used in Ruts et al. (2004) and De Deyne et al. (2008) and 3 scale-points for the 9- and 7- point-scales used in Hampton's French dataset and experiment 3, respectively. This corresponds to normalised IQRs greater or equal .26, .38 and .5, respectively. 554 SC typicality ratings have a high IQR according to this criterion and 190 of them have a dip p-value ≤.05 (19% of all rating distributions, 34% of rating distributions with high IQR). Table 15 shows how many SCs per category and dataset fulfil these criteria for multimodality. Ruts et al. (2004) and De Deyne et al. (2008) used almost the same categories, in both experiments the participants were native Dutch speakers and the percentage of multimodal distributions is similar for most categories except *fish*, *tools* and *vehicles*, where the De Deyne et al. (2008) data have fewer multimodal distributions. The means over all categories for their datasets are with 2.9 and 4.5 much smaller than those from the French dataset (9) and experiment 3 (7). The categories with highest mean of multimodal distributions are *science* (13, 54%), *insect* (9, 36%) and *sports* (8, 30%). *Profession* has only one multimodal distribution in both datasets in which it was used. In this category, there is no distinct typicality ordering as all SCs have a high mean typicality. *Reptiles* also has only one multimodal distribution. Upon inspection, I saw that while the category has a typicality ordering, there is low intersubjective agreement with almost uniform rating distributions, which means that there are no distinct modes.

Table 15: Number and percent of items with dip p-values ≤0.05 and with high IQR per category for all datasets.

dataset		amphibian	bird	fish	fruit	insect	mammals	musical instruments	profession	reptiles	sports	tools	vegetables	vehicles	clothing	kitchen utensils	weapons	furniture	science	mean	SD
Ruts et al.	number	3	2	6	5	8	3	4	1	1	7	6	6	6						4.5	2.3
(2004)	%	60%	7%	26%	17%	31%	10%	15%	3%	5%	23%	20%	20%	20%						20%	15%
De Deyne et al.	number	2	1	1	7	9	3	4	1	1		2	3	1	4	4	1			2.9	2.6
(2008)	%	40%	3%	4%	23%	35%	10%	15%	3%	5%		7%	10%	3%	14%	12%	5%			13%	13%
Hampton FR	number			7	5	10					9	12	11					7	13	9.0	2.6
(2017)	%			29%	21%	42%					38%	50%	46%					29%	54%	38%	11%
Experiment 3	number				5								9							7.0	2.8
(2018)	%				23%								41%							32%	13%
	mean	3	2	5	6	9	3	4	1	1	8	7	7	4	4	4	1	7	13		
	SD	0.7	0.7	3.2	1.0	1.0	0.0	0.0	0.0	0.0	1.4	5.0	3.5	3.5							

This analysis shows that the typicality ordering might in fact not be as intersubjectively stable as commonly assumed for the medium typicality level. Instead of showing high intersubjective agreement about the medium typicality of the SCs, the medium typicality means are the result of a high intersubjective variability. Means are severely distorted by outliers such that even a few ratings deviating from the dominant response can change the mean value drastically (see section 3.1.2). For multimodal distributions, means do not reflect the central tendency well. If 20 participants rate an SC as highly typical and 20 as highly untypical, the mean will reflect a medium typicality, while SD and IQR have their maximal values. Due to this, SD and IQR should be reported with mean typicality ratings. It is interesting that the multimodality is found in similar amounts in the different datasets, which indicates that the intersubjective disagreement for medium typicality levels is stable across contexts and might be due to different interpretations of what typicality is that are held in roughly equal amounts in different populations of participants. Barsalou (1987) also reports various results that indicate that typicality judgements change with different contexts and for the same participants over time.

3.3 Predictors of typicality

Two variables have often been assumed to be closely related to typicality: how often an SC is generated in productive frequency experiments and how familiar an SC is. Typical SCs are often generated in PF tasks and rated to be familiar. Section 3.3.1 compares the correlations between typicality and PF and section 3.3.2 the correlation between familiarity and typicality.

3.3.1 Productive frequency

Productive frequency (PF) measures how many participants gave a particular response, within a given set of participants. To collect it, participants are given a blank sheet with a category name on the top and asked to produce as many SCs as possible within a certain time (30 seconds in Battig and Montague (1969), 60 seconds in Hampton and Gardiner (1983), 15 seconds in Barsalou (1985), 30 seconds in Ruts et al. (2004), Schröder et al. (2012) gave no time limit).

Mervis et al. (1976) report rank order correlations between .48 and .74 of the Rosch (1975b) typicality data with PF data from Battig and Montague (1969) (.7 for fruit and .74 for vegetables). Hampton and Gardiner (1983) report a mean correlation of -.76[8] (-.87 for fruit and

[8] Pearson correlations between typicality and PF are negative if the highest typicality corresponds to the lowest scale point.

-.86 for vegetables) and that, in comparison, British and American PF data had lower correlations (.76) than British and American typicality ratings (0.85)[9], which they note is in correspondence with Rosch (1978)'s statement that "[t]ypicality depends more on the family resemblances among items, whereas associative frequency may be expected to reflect local differences in language use and item familiarity." (Hampton & Gardiner, 1983, p. 498). Barsalou (1985) reports a mean correlation of .39 for goal-derived categories and .55 for common taxonomic categories. For fruit, it is .72 and for vegetables .42. In the data provided by Ruts et al. (2004), the correlation is .67 for fruit and .59 for vegetables. In the data from De Deyne et al. (2008), the correlation is .73 for fruit and .68 for vegetables. In the dataset from Schröder et al. (2012), the correlation between PF and typicality is .61 for fruit and .56 for vegetables.

Despite showing a clear relationship, PF and typicality ratings are not identical. It is often assumed that PF is guided more by salience than typicality. Hampton (1997a) shows in two experiments that they are clearly distinct cognitive variables relevant to categorisation, where PF is based on associative retrieval and typicality on a similarity comparison (p. 637).

3.3.2 **Familiarity**

The variable familiarity is obtained like typicality: participants are asked to rate how familiar they are with SCs. It is collected in many typicality studies.

Hampton and Gardiner (1983) collected familiarity ratings in the same way as typicality on a scale from 1 (very familiar) to 5 (very unfamiliar) with the instruction "to rate the words according to how familiar you are with their meaning" (p. 494). They report a mean correlation of .54 with typicality

Schwanenflugel and Rey (1986) used the same instructions like Hampton and Gardiner (1983), except that they extended the scale to 7 points and reversed it, so that 1 means highly unfamiliar and 7 means highly familiar, and translated them to Spanish. They found that familiarity correlations between the English and Spanish speaking participants were lower (.48) than typicality correlations (.64). When controlling cultural familiarity, they found that there was a higher correlation of typicality ratings between English and Spanish speakers of .73 ($F(2,674) = 389.47$, $p<0.001$) and conclude that "discrepancies in familiarity do play a role in reducing intercultural overlap in typicality." (p. 155).

De Deyne et al. (2008) operationalised familiarity as estimations of how often participants had used the word by which the concept is expressed. "The participants were asked to indicate, for every item in the list, how familiar they were with the item. The participants used a Likert-type rating scale, ranging from 1 to 5. They were instructed that a 1 meant that *they had never seen, heard, or used the word before*; a 2 meant *that they had seen, heard, or used it just once or twice*; a 3 meant that *they had seen, heard, or used it sometimes*; a 4 meant that *they had seen, heard, or used it often*; and a 5 meant that *they had seen, heard, or used it very often*." (p. 1036). The correlation with typicality is for fruit .91 and for vegetables .19.

Schröder et al. (2012) measured familiarity by asking subjects to "estimate the degree to which they thought about or came in contact with a concept", using a 5-point scale ranging from 1 (very unfamiliar) to 5 (very familiar). Care was taken to make sure that the estimate had been attributed to the concept itself and not the word." (p. 384). They report a mean correlation over

[9] They compared the correlations between 12 resp. 8 categories.

all categories of -.59[10] between typicality and familiarity. For *fruit* it is -.88 and for *vegetable* -.75.

Moreno-Martínez et al. (2014) formulated the task as follows:

> "Your task will be to judge the 'familiarity' of each word – that is, the degree to which you come in contact with or think about the concept in your daily life. Please use the following 5-point scale in order to judge how familiar or common – or unfamiliar or uncommon – an object is for you. A '1' means that this is a very unfamiliar object. A '5' means that this is a very familiar object." (Moreno-Martínez et al., 2014, pp. 1094–1095).

They report a mean correlation of .85 with typicality ($p<.01$). For *fruit* it is .89 and for *vegetable* .86.

Like PF, familiarity has medium to high correlation with typicality. Also like PF, typicality is not reducible to typicality, as the correlations are not perfect and controlling for familiarity actually raised typicality correlations between English and Spanish speakers and correlations between British and American familiarity ratings were lower than those of the typicality ratings. Furthermore, the studies used different interpretations of familiarity. For some, it reflects the familiarity with the category designated by the word and for some the familiarity with the word.

3.4 **Discussion**

The results from this chapter have some implications on how typicality orderings should be interpreted. What can be seen is that participants – independent from their culture, the experimental instructions and their native language – agree on which SCs have a high typicality for the fruit category. For the vegetable category, the correlations between studies were lower for different cultural and language groups, but high within these groups. There is thus, at least partial, agreement about which SCs are on the high typicality level. On the individual study level, there are unimodal distributions and high intersubjective agreement. The same amount of agreement was not seen for medium and low typicality levels. It begins with different criteria for the selection of SCs which led in most cases to many more SCs on the high typicality level. The different selection criteria are reflected in different ways to interpret the lowest scale point and to provide opt-outs. There are fewer SCs on the lower typicality levels in each study and even fewer of them are used in many studies and can be compared. Furthermore, their mean ratings are less clearly interpretable because the meaning varies between scales with differing lower scale points. On the level of the individual studies, medium mean typicality is often due to multimodal or uniform distributions for which the mean is no suitable measure of central tendency. The French typicality data were the only ones which included clear non-members and here it was seen that there is intersubjective agreement on the lowest typicality level, if minimal typicality is taken to mean non-membership, which most authors seem to agree to[11].

Due to these uncertainties, the reference to an intersubjective typicality *ordering* of common-sense categories seems to be not supported by the data, at least not in the sense of small mean differences reflecting small typicality differences. Three unique distribution patterns suggest the distinction of three primary typicality levels:

[10] The correlation is negative because typicality and familiarity were measured in different directions.

[11] There are counterexamples to this: "For example, bats were moderately typical of birds, but were very unlikely to be included in the category. Dolphins and whales had a similar relation to the category fish." (Hampton and Passanisi (2016, p. 507)).

1) highly typical ones with high intersubjective stability and a unimodal rating distribution,
2) medium typical ones with low intersubjective stability and multimodal rating distributions,
3) very untypical ones or non-members with high intersubjective stability and a unimodal rating distribution.

In this view, apples are not more or less typical for fruit than pears, both are highly typical. Apples and pears are however still more typical than pumpkins, which are medium typical, and apples, pears and pumpkins are still more typical than milk or carrots, which are non-members or atypical depending on the interpretation of the lower end of the scale. Assuming that only these three levels have cognitive reality introduces the problem of second-order vagueness: how exactly are the levels differentiated? What percentage of responses counts as unimodal?

What can certainly be concluded is that there is a typicality ordering that is subject to high levels of measurement error and subjective unreliability – the closer two items are on the scale, the harder it becomes to determine their order. It would, however, always be possible to fix the concrete order with a large sample size. For parametric models that predict typicality with typicality-contributing properties, this means that the concrete rank order or mean values, in particular when they were determined from small sample sizes, should be taken with a grain of salt.[12]

[12] These thoughts are from James Hampton (P.C.).

4 Formal models of prototype theory

This chapter presents theoretical and practical considerations for the construction of a formal model of prototype theory, which is one of the goals of this thesis. I present a formalisation of prototype theory's main hypotheses in the first section (4.1) that closes with the general formula

$$\mathrm{typ}(\mathrm{SC}, \mathrm{C}) = \mathrm{f}\Big(\mathrm{Sim}\left(\mathrm{SC}, \mathrm{C}|\mathrm{P_{typ}}(\mathrm{C})\right)\Big),$$

that is, the typicality typ of an SC for a C is a function f of the similarity Sim of SC to C in terms of typical properties $\mathrm{P_{typ}}$ of C. I then successively discuss each of the formula's components. First, the identification of $\mathrm{P_{typ}}$ in section 4.2. Then, the representation of quantified $\mathrm{P_{typ}}$ in SC-representations and C-representations in section 4.3. The last section 4.4 presents possibilities to determine Sim for these representations.

4.1 Formalisation

The empirical finding that common-sense categories (Cs) have an intersubjectively stable typicality ordering of their SCs[13] is described in the *typicality ordering observation* (TO):

(TO) For each subcategory $\mathrm{SC_i}$ of C on an abstraction level lower than C exists an intersubjectively stable typicality level $\mathrm{t_i}$ of $\mathrm{SC_i}$ for C, $\mathrm{typ}(\mathrm{SC_i}, \mathrm{C}) = \mathrm{t_i}$. The typicality ordering of C corresponds to the mathematical structure $\langle \mathrm{SC_i}, \leq_{\mathrm{t_i}} \rangle$, a partially ordered set of all SCs' typicality levels according to the smaller or equal relation $\leq_{t_i}$.

In prototype theory, (TO) is explained to be the result of a comparison of SCs to a prototype representation of C, which is the *prototype theory hypothesis* (PTH). The prototype representation is either equated to the most typical SC or exemplar (PTH-E) or to a summary representation of the typical properties of C (PTH-S):

(PTH) The typicality ordering of C arises from the assessment of an SC's similarity to C's prototype PT(C):

$$\mathrm{typ}(\mathrm{SC}, \mathrm{C}) = \mathrm{f}\Big(\mathrm{Sim}\big(\mathrm{SC}, \mathrm{C}|\mathrm{PT}(\mathrm{C})\big)\Big).$$

PT(C) is specified either as:

(PTH-E) the most typical SC (prototypical exemplar prototype), or

$$\mathrm{PT}(\mathrm{C}) = \mathrm{argmax}_{\mathrm{SC_i} \in \{\mathrm{SC_1}, \ldots, \mathrm{SC_n}\}}(\mathrm{typ}(\mathrm{SC_i}, \mathrm{C}), \text{ or}$$

(PTH-S) a summary representation of the typical properties of C, $\mathrm{P_{typ}}(\mathrm{C})$, (summary representation prototype):

$$\mathrm{PT}(\mathrm{C}) = \mathrm{P_{typ}}(\mathrm{C}).$$

For (PTH-E), only the similarity of each SC to the one most typical SC is important. For (PTH-S), the prototypical properties of C are important. They are usually taken to be those that are present in many (but not necessarily all) SCs. The highly typical SCs share many properties

[13] Chapter 3 raised doubts about the significance of small mean differences for the typicality ordering. We assume that the concrete rank orders can be determined with a large enough sample size.

between each other, while the untypical ones share few or none. This way of thinking is called the *family resemblance hypothesis* (FRH) and goes back to Wittgenstein (1953)'s natural language philosophy:

(FRH) SCs share properties between each other. Those properties that are found in many (but not necessarily all) SCs are the typical properties of C, $P_{typ}(C)$.

(FRH) does not state in how many SCs the properties have to be found to count as typical properties. In Rosch and Mervis (1975), all properties that are present in at least 2 SCs are counted as typical. Their typicality contribution rises with the number of SCs in which they are present. Our model also determines typicality contribution from the frequency of properties in SCs and Cs, but not exclusively. It will be presented in chapter 5.

4.2 **Properties**

It is not trivial to identify relevant properties out of the many that each object has. As noted in Goodman's (1972, p. 443) famous critique of similarity, it does not matter which two objects are compared, they all have at least one property in common. Everything can be argued to be similar to everything else. For example, an umbrella and a ghost are similar in sharing the property *can be transparent*. Giving similarity a deeper, more useful meaning then depends on identifying relevant dimensions of comparison. Goodman deems this impossible, because "importance is a highly volatile matter, varying with every shift of context and interest, and quite incapable of supporting the fixed distinctions that philosophers so often seek to rest upon it." (ibid., p. 444). That similarity is context-dependent was empirically confirmed in Tversky (1977). But this does not mean a) that it is impossible to identify relevant dimensions for specific context types and b) that there is no default context, used in the absence of specific information. Typicality ratings seem to introduce such a context and typical properties give an indication of the relevant dimensions. In absence of more specific information, it is plausible that humans apply this context as default context when interpreting statements containing category concepts. Then, hearing the sentence "there was a fruit on the table" would allow me to infer that this object has typical fruit properties, like that it probably tasted sweet and grew on a tree. As mentioned before, Barsalou (1985) found stable typicality orderings for ad hoc categories like *things to take from one's home during a fire*, indicating that people agree on the relevant dimensions even in spontaneously (ad hoc) generated contexts.

My goal is to find a cognitively plausible representation of the mental representations of categories. In this context, psychological experiments are the best source to identify relevant category properties. Property generation data are the most popular format for the investigation of conceptual structure. The generated properties are however not necessarily important in similarity judgements and rather a first indication of possible candidates for typical properties. A subsequent mechanism that identifies their typicality-contribution is required.

Most of the time, category properties are derived from pooling the properties that were generated for category members. Both Rosch and Mervis (1975) and Smith et al. (1988) used this procedure and used these properties for their typicality predictions (see section 8.1 resp. 0 for an in-depth description of their procedure). Tversky (1977) used a similar procedure to calculate the similarity between category members.

Two exceptions are found in the literature. De Deyne et al. (2008), who collected properties for categories and category members, and Hampton (1979), who asked 32 participants to list

properties of categories in interviews that aimed at generating as many properties as possible. The list of all generated properties per category was then pooled, all properties mentioned by fewer than 5 participants excluded and then analysed in terms of PF and importance ratings by 16 judges. All properties listed by fewer than 8 participants which had at the same time a low importance rating were excluded, resulting in 8 to 16 properties per category. Table 16 shows exemplarily the properties he identified for the fruit category. Hampton analysed in how far the presence of defining and characteristic properties predicts category membership[14]. Defining properties are those present in all SCs that are clear members and characteristic properties are present in more typical than atypical SCs. A property was regarded to be present on the basis of the average judgment of 4 judges in the interval [-8, 8], who rated applicability on a 5-point-scale (+2 = definitely yes, 0 = uncertain, -2 = definitely no). This can be called a gradual application score (GAS). He found that while the amount of both characteristic and defining properties held by an SC correlates positively with its category membership score ($\tau \in (.61,.78)$, $p<.001$, mean = .622) and weighting them with the GAS raised the correlation (mean = .674 (+.052)), weighting them with the definingness of properties, either as rated importance or PF, did not (mean = .643 for median ranked importance and mean = .660 for PF). (pp. 451-452).

Table 16: Properties generated for fruit reported in Hampton (1979) (p. 459).

No.	Feature	Defining or characteristic	Production frequency
1	is a plant, organic, vegetation	D	31
2	Is edible, is eaten	D	30
3	Contains seeds	C	27
4	Grows above ground, on bushes or trees	C	26
5	Is juicy, thirst quenching	C	17
6	Is brightly coloured	-	16
7	Is sweet	C	15
8	Has an outer layer of skin or peel	-	13
9	Is round	C	9
10	Is eaten as a dessert, snack, or on ist own	C	8
11	Is a protection for seeds	C	7

Two different ways of collecting property data are interesting and promising: large-scale word association data as collected and studied in the ongoing Small World of Words (SWOW) project presented in De Deyne et al. (2018) and properties derived from functional divisions in the brain as studied in Binder et al. (2016).

The SWOW project is an ongoing collection of associations for a very large number of words. It has a very large pool of participants: in 2018 they report 83,863 participants who generated 3,665,100 responses for 12,217 words. Figure 25 shows a visualisation available from the SWOW for fruit and vegetables in a one-hop network and Table 17 how an attribute-value-structure can be derived from their data format. For this, all first associations for fruit SCs were extracted and then their frequency within these SCs was computed. The table shows the most frequent ones. A comparison with the attribute-value-structure that I derived from PF data and

[14] Rated on a scale from 1 (definite category member) over 4 (unable to decide) to 7 (totally unrelated) by 32 participants, then summed and linearly transformed to a 100-point-scale with 100 signifying absolute intersubjective agreement about membership and 0 signifying absolute intersubjective agreement about non-membership. This scale was significantly correlated both with PF from Battig and Montague (1969) ($\tau \in [.39,.69], p < .005$) and typicality from Rosch (1975) ($\tau \in [.65,.80], p < .001$) (p. 445-446).

probability ratings (Table 48) reveals that the two ways lead to very similar structures. Generating attributes and values from association data requires however more interpretation than generating them from properties, because associations are single words that have to be interpreted to form a property. Their advantage is that they are readily available and require no separate data collection.

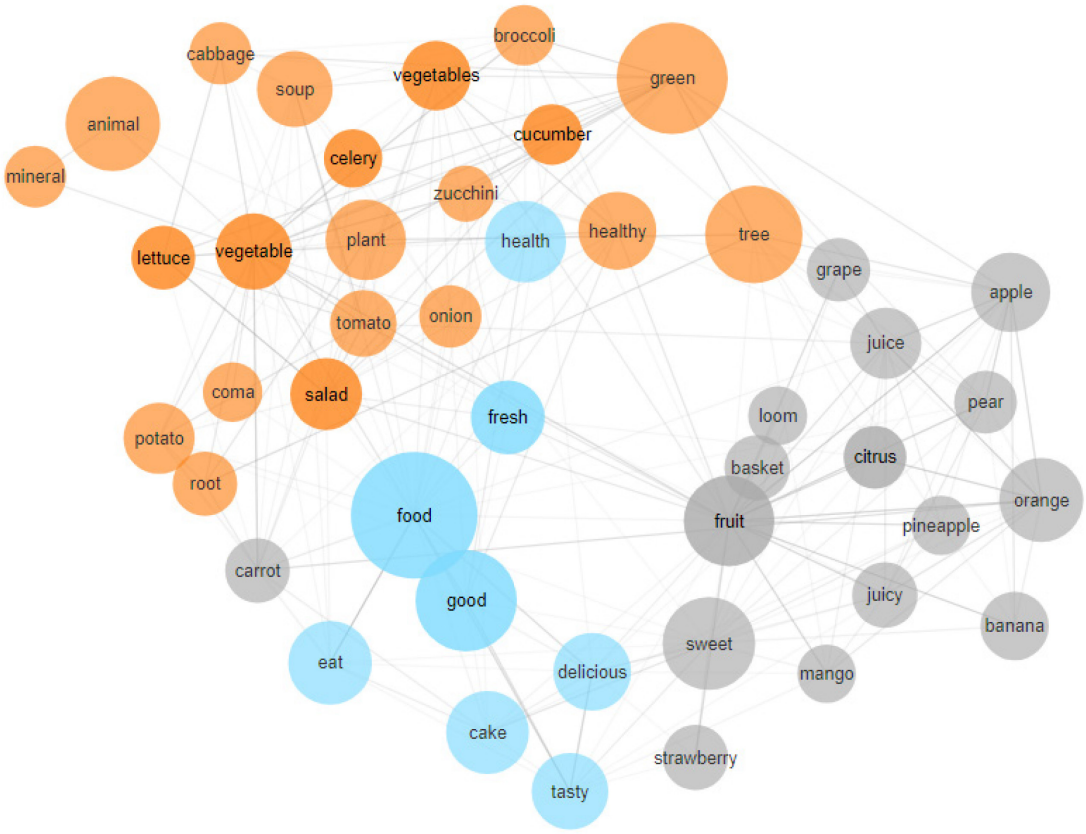

Figure 25: Visualisation of a one-hop network generated on https://smallworldofwords.org/en/project/visualize for fruit and vegetable.

Table 17: Possible way to derive an attribute-value-structure from the SWOW data.

response	frequency	attribute	value	property
fruit	46	supercategory		
juice	29	CONSUMED-IN	Juice	will be made into juice
sweet	28	TASTE	Sweet	tastes sweet
food	27	supercategory		
tree	23	HOW-GROWN	On a tree	grew on a tree
orange	22	COLOUR	Orange	is orange
apple	21	subcategory		
red	20	COLOUR	Red	is red
green	18	COLOUR	Green	is green
delicious	18	TASTE	Delicious	tastes delicious
pie	17	CONSUMED-IN	Pie	will be eaten in pie
vegetable	14	supercategory		
yellow	14	COLOUR	Yellow	is yellow
yummy	13	TASTE	Delicious	tastes delicious
nut	12	supercategory		
jam	12	CONSUMED-IN	Jam	will be eaten in jam
purple	11	COLOUR	Purple	is purple
juicy	11	JUICINESS	Juicy	is juicy
eat	11	action		
sour	10	TASTE	Sour	tastes sour
color	10	supercategory		
yum	10	TASTE	Delicious	tastes delicious
tasty	10	TASTE	Delicious	tastes delicious

Binder et al. (2016) present results from letting participants rate how far certain words are associated with certain qualities that are known to involve different brain regions when processed like *pain*, *lower-limb activity*, *audition* and *vision*. Table 18 shows how a partial attribute-value-structure can be derived from their data. The structure is partial, because it involves many attributes for which the values have to be determined afterwards – fruit SCs were for example rated to be highly associated with a characteristic taste and colour. This indicates, in accordance with the data from the SWOW and my analysis, that these are important attributes for fruit. Which tastes and which colours are characteristic would have to be determined in a next step if this data were to be used. Other components, like visibility or associated body part, are too general to characterise fruit, because as property of the supercategory *food*, they apply to all SCs.

Table 18: Partial attribute-value-structure for fruit derived from Binder et al. (2016).

component	mean rating	attribute	value	property
Taste	5.54	TASTE	?	has a characteristic taste
Color	5.30	COLOUR	?	has a characteristic colour
Shape	4.92	SHAPE	?	has a characteristic shape
Vision	4.77	VISIBILITY	High	something that you can easily see
Weight	4.27	WEIGHT	?	light or heavy in weight
Small	4.16	SIZE	small	is small
Practice	4.14	USE-EXPERIENCE	Yes	a physical object you have experience using
Smell	4.09	SMELL	?	has a characteristic smell
Texture	4.01	TEXTURE	?	has a smooth or rough texture
Head	3.93	BODY-PART	Head	associated with actions using the head
Pleasant	3.73	EMOTION	Pleasant	something that you find pleasant

While both the data from the SWOW and from the brain component analysis require some additional processing before they can be used as property input, their use has some advantages as well. The SWOW project offers a huge database of word associations for many different concepts. A good procedure to incorporate their data for a subsequent prototype analysis would open the possibility to broadly extend and verify results with small effort. The brain component analysis offers a natural way of identifying and restricting the attributes and values used in representation, which is otherwise a subjective component in the models and therefore difficult to validate externally.

It has been indicated in many experiments and prominently argued by Lawrence Barsalou and colleagues (e.g., Barsalou et al., 2003, Barsalou, 2008) that conceptual processing, as a part of semantic memory and language processing in general, does not use amodal symbols, an abstract representation format that is distinct from other representations, but is instead "grounded" in the modality-specific systems of the brain. Rather than "translating" experiences into an abstract language of thought, the brain's modal systems are directly employed in conceptual processing. This observation is of interest because it can explain where the properties generated in generation tasks "come from". We have evolved to perceive the world in a certain way and this way is determining which properties we observe. A theoretical way to capture the naturalness of attributes and properties has been proposed for conceptual spaces (e.g., Douven & Gärdenfors, 2020). These results are interesting for the present research because they restrict the kind and number of (natural) attributes.

4.3 Representation and quantification

The typicality contribution of properties has so far been quantified in previous work either by determining how many SCs have the property (the property's applicability) or by counting how often the property was generated by participants for each SC (the property's number of votes).

Different ways to determine a property's applicability with application scores (AS) are found in the literature. Rosch and Mervis (1975) use a binary conception of applicability in which the AS is 1 if both judges j_1, j_2 agree that the property applies and 0 if they agree that it does not apply or if they disagree:

$$AS_{R\&M}(P_k) = \begin{cases} 1, & \text{if } j_1 = j_2 = 1 \\ 0, & \text{if } j_1 = j_2 = 0 \text{ or } j_1 \neq j_2 \end{cases}.$$

The property ASs are summed over all SCs and this number, the property's family resemblance score, determines the contribution of the property to the typicality of each SC that has them. Figure 26 shows an example. The property *is sweet* has received an applicability judgement of 1 by both judges for all three SCs and therefore has a property family resemblance score (FR score) of 3, while the property *is blue* applies only to one SC, *blueberry*, and receives therefore the weight 1. The list of the property family resemblance scores on the right is the quantified category or prototype representation in their model. Each SC's list of application scores is the SC representation.

property	apple	blueberry	pear
is sweet	1	1	1
is round	1	1	0
is blue	0	1	0
is green	1	0	1

property	FR(P)
is sweet	3
is round	2
is blue	1
is green	2

Figure 26: Example for application scores in Rosch and Mervis (1975).

Hampton (1979) collected applicability ratings between -2 (definitely not) and 2 (definitely yes) from four participants and calculated a graded application score which corresponds to the sum of all participants' ratings with values between -8 and 8:

$$\mathrm{GAS}_1(\mathrm{P_k}) = \sum_{\mathrm{i}=1}^{4} \mathrm{j_i}\,, \mathrm{j_i} \in \{-2,-1,0,1,2\}.$$

Djalal et al. (2017) also propose a graded application score which ranges between 0 and 4 depending on how many of 4 participants agreed to the applicability of the property:

$$\mathrm{GAS}_2(\mathrm{P_k}) = \sum_{\mathrm{i}=1}^{4} \mathrm{j_i}\,, \mathrm{j_i} \in \{0,1\}.$$

Determining the applicability of properties to SCs is essential to successfully represent them. It is however a vague notion which can be seen in the variations of the application scores proposed in the literature – is a property applicable when it applies to some of the objects categorised as SC, if it applies to most or does it have to apply to all objects? While Rosch and Mervis systematically excluded all cases in which the judges disagreed, Hampton and Djalal et al. incorporate their disagreement into the score. For example, certainly not all blueberries taste sweet, unripe ones tend to be sour. Some judges might say the property still applies to blueberries in general and some may say that it does not or on Hampton's scale that it applies partly with a rating of 1. This would result in different application scores in each version.

Instead of application scores, some authors have used the properties' productive frequency (PF) as a measure of property applicability. But this has disadvantages as well, as PF seems to be mostly guided by salience. While being an important variable, salience does not equal importance for typicality. This was noted in Malt and Smith (1984):

> "Frequency of listing did not seem to be an adequate measure, since a number of spurious factors appeared to influence the production frequency of a property for a given member. 'Has wings', for instance, might be listed with high frequency for 'robin' but less frequently for 'penguin' because other properties were relatively more salient for penguins." (Malt & Smith, 1984, p. 256)

When the single properties' contributions are quantified, a representation format for Cs and SCs is required to compare them. As discussed in section 1.2, representing concepts in feature lists is straightforward: the properties, together with their corresponding weights, are collected in a

list like on the right side of Figure 26. If, however, a deeper underlying structure is assumed and the concept is supposed to be formalised in a quantified prototype frame (QPF), an additional step of identifying the attribute-value-pairs underlying the properties is required.

Smith et al. (1988) asked two judges to determine intuitively which properties belong to the same attributes. Another way would be to base the attribute selection on the domains identified in Binder et al. (2016) or on the classification proposed in Wu and Barsalou (2009).

Cohen and Murphy (1984) argue that, while set-theory is a good tool to model the structure of the environment, it is insufficient for modelling conceptual structure. They propose an A-V-structure for categories instead. They call attributes roles that are "filled by feature lists, called 'values'" (p.45), and each role's range of values is restricted. The relation between Cs and SCs is defined as an "organization of concepts into a lattice or taxonomy supporting inheritance of roles" (p.45) and "values for a given role may be ordered by typicality" (p.46). "Objects that have the most typical role-values will be considered 'good' members; objects that have atypical or unacceptable values will be considered 'borderline' or 'poor' members." (p.48). This is the basic idea behind QPFs.

Hampton (1993) also notes that the most plausible representation of C prototypes are QPFs:

> "A prototype concept is constituted by a *set of attributes with associated values* […], each with a particular weight corresponding to its 'definingness' or contribution to the concept's definition. More generally, the weight of an attribute may be thought of as a distribution of weights defined across a range of values for the attribute. […] Information will thus be included on the permitted variability across category members in the value of any attribute. Weights will also vary between attributes – the relative weight of an attribute can be defined as the weight of the modal value of the weight-value distribution." (Hampton, 1993, p. 73)

An advanced proposal of a QPF is presented in Smith et al. (1988). Figure 27 shows the components of the category prototype and representations of SCs (which they refer to as I for "instances"). Their C representation has three components: the A-V-structure, a measure of attribute diagnosticity and a measure of value salience:

> "Subsequent work has shown that the contents of a prototype must include far more than a list of properties. For one thing, we need to decompose the notion of a property into two components: *attribute* and *value*. […] A prototype also includes some indication of the salience of each relevant value. […] Finally, a prototype may also include some indication of the diagnosticity of each attribute, that is, a measure of how useful the attribute is in discriminating instances of the concept from instances of contrasting concepts." (Smith et al., 1988, p. 487)

Their SC representation has the same structure as the prototype, but they are instantiated, i.e., the salience weights are shifted to one single value. In an earlier version, they call their proposal a frame explicitly: "This kind of representation is essentially a frame (e.g., Minsky, 1975; Winston & Horn, 1981), with *attributes* being *slot-names*, *features* being *values*, and *most-likely features* being *default values*." (Smith & Osherson, 1984, p. 353). Different from this earlier specification, they added salience weights to all values instead of only marking the most probable one.

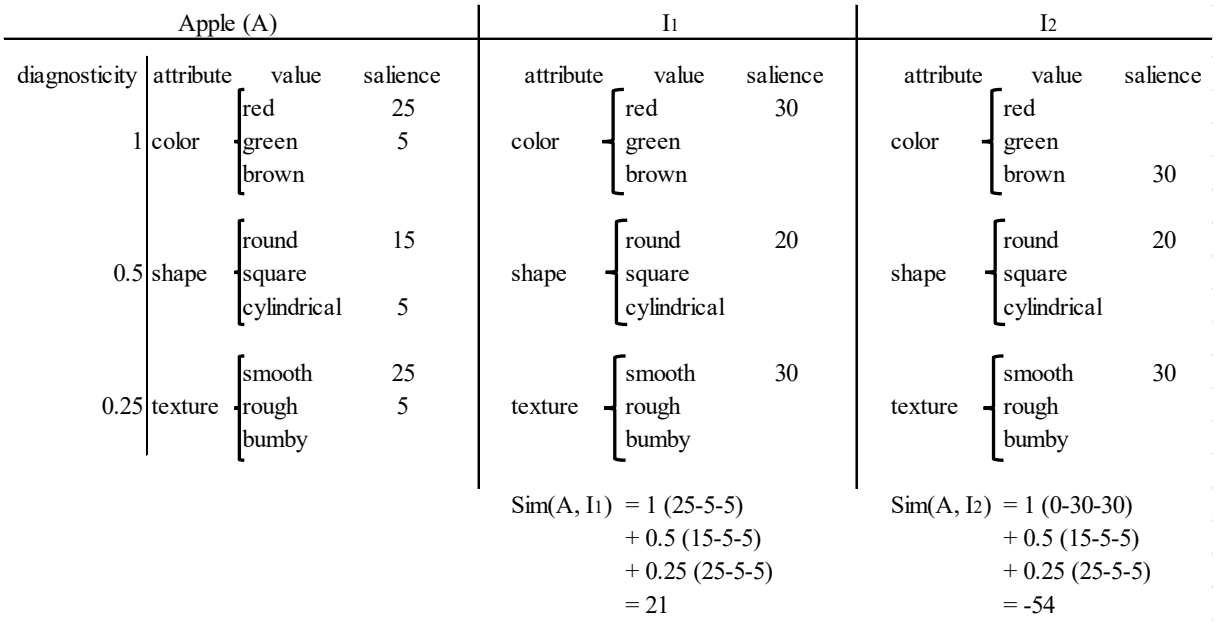

Figure 27: Figure 1 from (Smith et al., 1988, p. 490): "Illustration of attribute-value representations for a prototype (apple) and relevant instances (a red apple (I_1) and a brown apple(I_2)); beneath each instance representation is the similarity between the instance and prototype."

Sutton (2017 and unpublished manuscript) proposes to represent prototypes in Bayesian networks, as illustrated in Figure 28. Each concept (root node) is characterised by attributes and each attribute value is quantified with a probability between 0 and 1. For example, the frame for *pet* has a TEXTURE attribute with 3 values. The most probable one is Furry with 0.8 probability, the other two, Scaly and Feathered, are improbable with 0.1 probability each. His proposal is very close to what I develop in chapter 5.

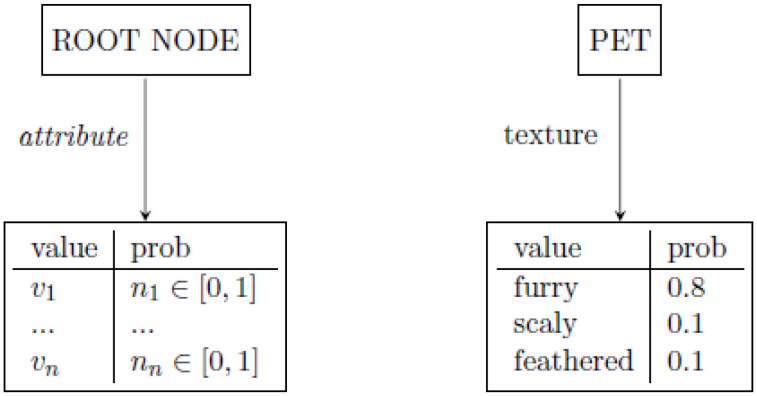

Figure 28: Figure 1 from Sutton (u.ms.), (Partial) Frame Schema (left) and Partial PET Frame (right).

In the same spirit, in Schuster et al. (2020), Peter Sutton, Corina Strößner, Henk Zeevat and myself discuss properties of stochastic frames and argue that attribute values can be represented as joint probability distributions, which makes them ideal tools for explaining constraints, vagueness, prototypes and lexical ambiguity. An interesting example for a stochastic frame from Strößner (2020b) is in Figure 29. It shows two related attributes of the *bird* concept: FOOT STRUCTURE and LOCOMOTION. All attributes are quantified with a subjective probability distribution reflecting for example that people estimate 80% of birds to have clawed feet and 20% of birds to have webbed feet. Furthermore, that out of the 50% of birds that are seen in

motion, 75% fly and 25% either walk or swim. While these values are not mutually exclusive for birds in general, because most birds can fly, walk and swim, they are mutually exclusive in a single observation. Additionally, constraints in terms of conditional probabilities are embedded, one reflecting that flying happens at a fast speed, $\Pr(\text{fast}|\text{flying}) = 1$ and one reflecting that 75% of birds with webbed feet swim, $\Pr(\text{swimming}|\text{webbed}) = .75$.

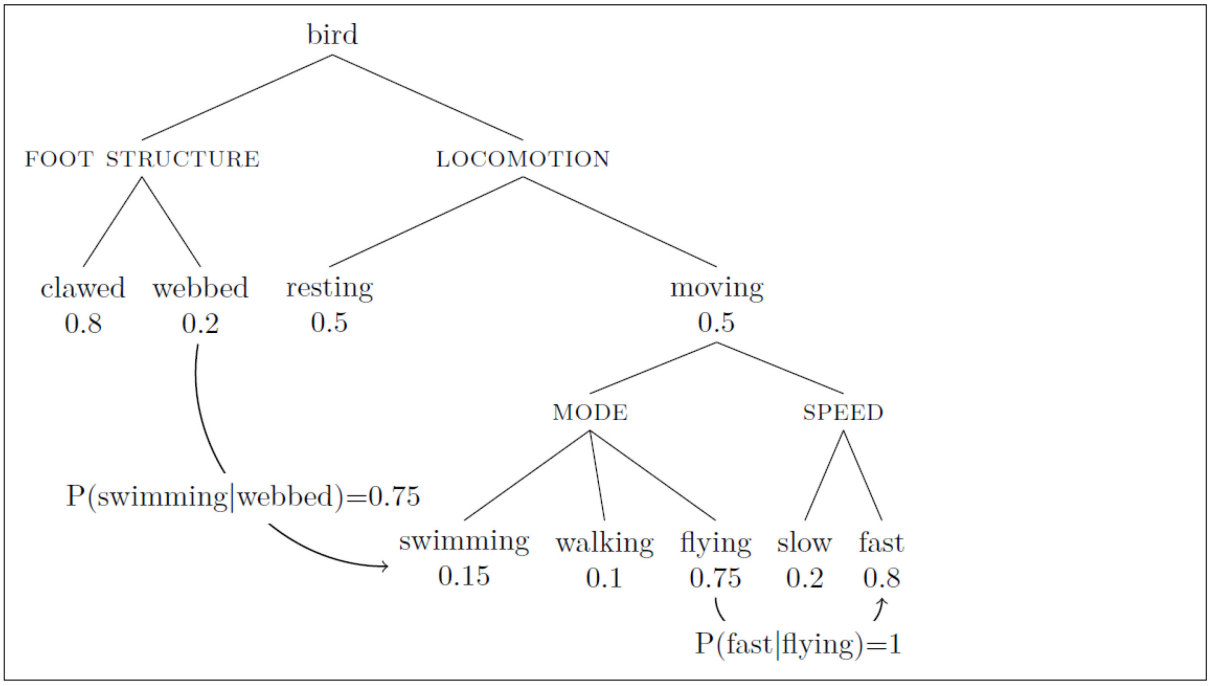

Figure 29: A partial representation of bird with a hierarchical structure and constraints between different levels (Strößner, 2020b, p. 698).

QPFs have been used theoretically and empirically to uncover the importance of different components of prototype concepts. They offer a fine-grained representation which readily incorporates quantified information of several kinds. The concrete quantifications proposed in the literature, applicability and PF, are, as I have argued, not ideal. In chapter 5, I present arguments to use a specific kind of probability information in QPFs, like in the proposals by Sutton and Strößner.

4.4 Similarity

The last section discussed ways to identify properties that are relevant in similarity determination and to quantify their contribution. When the contributions of all individual properties to similarity have been determined and parametric representations of Cs and SCs incorporating the contributions were selected, the individual contributions have to be aggregated in order to quantify the overall similarity of an SC to the C prototype. The computed overall similarity should then be highly correlated to the SCs' typicality ratings for C according to the family resemblance hypothesis.

If C and SCs are represented as property lists, like in Rosch and Mervis (1975) or Tversky (1977), the individual contributions are directly used as input of the aggregation function. If they are represented in frames, the properties are structured in attributes and an intermediary step can assign weights to attributes, like in Smith et al. (1988). All three proposals will be discussed in what follows.

Mathematically, concepts can be represented as sets of properties or as points in an n-dimensional geometrical space. Feature lists and frames employ set theory and a developed model of the second kind are conceptual spaces as introduced in Gärdenfors (2000). Strößner (2020b) discusses a possible integration of conceptual spaces into frames. Then it is possible to use methods from vector geometry for the similarity aggregation. I discuss similarity measures for sets and vectors in what follows.

Assume that a set of typical properties $PT = \{p_1, \dots, p_n\}$ was identified and that $P = \{(p_1, q_1), \dots, (p_n, q_n)\}$ and $SC_i = \{(p_1, q_{i1}), \dots, (p_n, q_{in})\}$ are sets of ordered pairs which contain these properties and a quantification q for each property for the category prototype P and the i SCs that are analysed. Four measures are potentially relevant for the comparison of the SCs with P, as summarised in Lesot et al. (2009):

- a: the amount shared between P and SC, which can be the intersection of the two sets, $(P \cap SC_i)$, if the quantification is binary, i.e., based on a list (with "1" for "in-list" and "0" for "not-in-list") or the minimal value, $\sum_j \min(q_j, q_{ij})$, if it is a continuous measure,
- b: the amount uniquely present in the prototype, which can be the set difference, $(P - SC_i)$ for binary quantifications, or the sum of the differences, $\sum_j q_j - q_{ij}$, if it is a continuous measure,
- c: the amount uniquely present in the SC, analogous to b either $(SC_i - P)$ or $\sum_j q_{ij} - q_j$, and
- d: the amount that is absent in both, which corresponds to the complement if the quantifications are binary, $(\overline{P \cap SC_i})$ and to $\sum_j (\max(q_j) - q_j) + (\max(q_j) - q_{ij})$ for continuous measures and is thus only defined if a maximal value is defined.

Rosch and Mervis (1975) only use measure a in their calculations and a weighted sum as aggregation function, as illustrated in Figure 30. Each property AS is weighted by the property's FR score and the results are summed up. The resulting value is the SC's FR score. The example predicts a typicality ordering of apple>blueberry>pear. It is noteworthy that the FR score of an SC has no upper limit and depends on the amount of SCs and properties that are considered in its calculation. It can therefore only predict the rank order of the typicality ordering and no concrete values.[15] An in-depth discussion of their model is in section 8.1.

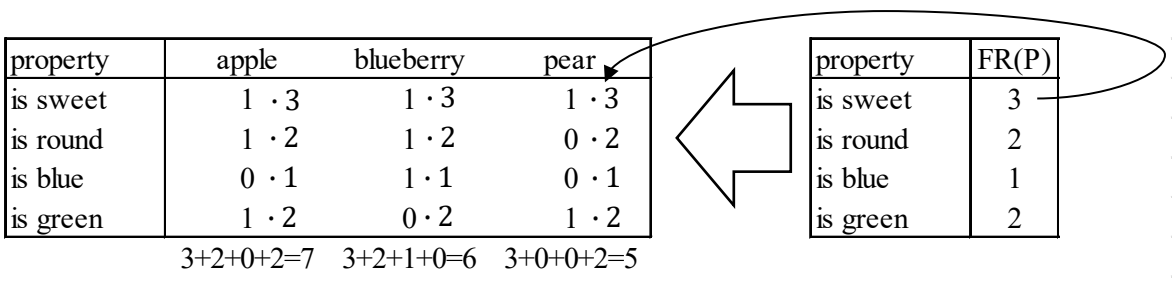

Figure 30: Calculation of the family resemblance score for three hypothetical SCs and four hypothetical properties.

[15] It might be possible to scale the FR property and then try to generate a function to create an interval scale of typicality. (James Hampton, P.C.)

Tversky (1977) introduces the contrast model, which is based on the thought that similarity is not only a matter of how many properties two objects share, but also of how many unique properties they have, thus b and c. He generalises all similarity measures possible with a, b and c as his contrast model,

$$\mathrm{Sim}(P, SC_i) = \theta f a - \alpha f b - \beta f c, \qquad \theta, \alpha, \beta \geq 0, (\mathrm{p.}\,332)$$

where θ, α, β are parameters which determine the importance of each component and f is an interval scale on which the properties are measured. Another formulation is the ratio model in which similarity is normalised to values between 0 and 1:

$$\mathrm{Sim}(p, sc_i) = \frac{a}{a + \alpha \cdot b + \beta \cdot c}, \qquad \alpha, \beta \geq 0 \ \ (\mathrm{p.}\,333).$$

The parameters α and β determine how much weight the unique properties of P and SC receive. Resulting similarity measures are only symmetrical if $\alpha = \beta$ or $b = c$, i.e., either if the task is non-directional or the objects are equal in the measured dimensions.

The results from Tversky (1977, pp. 338-39) are in Table 19. He found that the mean rated similarity of SCs from the category vehicle is positively correlated with similarity calculated as the sum of a, the common properties, (.68) and with the contrast model without parameters, a-b-c, (.72). These correlations rise to .84 and .87, respectively, if not only the binary applicability information (yes or no) is considered, but also the number of participants who listed the properties (the properties' PF). The distinct properties, b+c, are negatively correlated with SC similarity (-.34) and this correlation is also higher (-.64) when PF information is incorporated into the equation. Noting that properties should be quantified to account for the relative salience of each property and for the fact that PF does not necessarily correspond to it, the additive tree procedure (Sattah, Tversky, 1977) was used to determine the importance of each property in SC similarity. Using these quantifications raised the correlation to .94 for the contrast model.

Table 19: Reported correlations between common properties, distinctive properties as well as the contrast model with similarity between objects for 3 property measures: shared properties, productive frequency and additive tree procedure from Tversky (1977, pp. 338-39).

property measure	a	b+c	a-b-c
shared properties	.68	-.36	.72
productive frequency	.84	-.64	.87
additive tree procedure			.94

A frame-adapted version of the contrast model is presented in Smith et al. (1988). They propose to determine similarity to the prototype as diagnosticity-weighted sum:

$$\mathrm{Sim}(P, SC) = \sum_i v_i \sum_j \left[\theta \min\left(n_{ij}(P), n_{ij}(SC)\right) \dot{-} \alpha\left(n_{ij}(P) - n_{ij}(SC)\right) \dot{-} \beta\left(n_{ij}(SC) - n_{ij}(P)\right)\right],$$

where i is the index for attributes and j is the index for values of the attributes, v_i is the diagnosticity of the attributes and n_{ij} is the number of votes that the jth value of the ith attribute received, i.e., the amount of people who generated the value in question. They found that their model can predict the typicality ordering of adjective-noun-combinations very well and, less perfectly, the typicality ordering of categories. Their model is discussed in detail in section 8.2.

Some measures utilise d and give it the same importance as a, which corresponds to the thought that having the same properties as well as having same properties absent contributes to similarity[16], for example in the simple matching measure (Lesot et al., 2009, p. 69):

$$\mathrm{Sim}(P, SC_i) = \frac{a + d}{a + b + c + d}.$$

Tversky (1977) famously argued against the formalisation of similarity as geometrical mathematical entities, showing empirically that participant's similarity ratings do not follow the distance axioms, identity of indiscernibles, symmetry, triangle inequality and non-negativity. For example, he presents data showing that similarity statements are directional: "x is like y" was rated differently than "y is like x", for example North Korea was rated to be more similar to China than China to North Korea. According to Tversky, this is due to x introducing a reference frame to which y is compared. Gärdenfors (2000, pp. 113–114) discusses Tversky's criticism and argues that it does not apply to his version of geometrical models of concepts, because his model emphasises the importance of the salience of dimensions. In different similarity ratings, different dimensions can be salient: "the conceptual space is 'stretched' along the attended dimensions of the comparison" (p.113) – if "y is like x" leads to a shift of attention compared to "x is like y", the results of the comparison are different. Similarity judgements are not necessarily symmetric in conceptual spaces.

If Cs and SCs are represented as vectors with the prototype $p = (q_1, \dots, q_n)$[17] and the subcategories $sc_i = (q_{i1}, \dots, q_{in})$, similarity is based on distance measures, derived from the Minkowski distance:

$$\mathrm{dis}(p, sc_i) = \left(\sum_j |p_j - sc_{ij}|^{\gamma} \right)^{\frac{1}{\gamma}}$$

The most common ones are the Manhattan distance with $\gamma = 1$, which corresponds to the sum of the projections of the line segments between the points onto the coordinate axes, and the Euclidean distance with $\gamma = 2$, which corresponds to the line segment connecting them in a coordinate system and in which the difference terms are quadratically weighted into the distance function. Similarity can then simply be defined as $\mathrm{Sim}(p, sc_i) = 1 - \mathrm{dis}(p, sc_i)$. Other measures use the exponential function, for example Estes (1994), referring to Nosofsky (1984) and Shepard (1987) and his earlier work:

$$s_{ij} = e^{-c\,dis_{ij}} \ (p. 70),$$

where c is a sensitivity parameter controlling the steepness of the exponential curve. The use of the exponential function is motivated by the assumption that similarity "declines as a negatively accelerated function of the distance" (ibid.).

The most promising similarity scales for frame representations employ a feature-matching procedure. Due to the metric nature of probability data, it would however be possible to apply a variation of the Minkowski distance on each value or attribute and identify a formula that

[16] Counting shared absent properties works only when the set of properties is closed. Otherwise, there is an infinite number of properties which are not true of any two concepts. (James Hampton, P.C.)

[17] This is a simple vector that can only account for prototypes with one single dimension. In n-dimensional vector spaces, each attribute is represented with one vector.

unites them plausibly into a single coefficient, like a sum, average or weighted average. This option is not investigated in this thesis.

5 Probabilistic prototype frames

Based on the insights from the foregoing chapters, this chapter presents the conceptual representation format investigated in the remainder of this thesis: probabilistic prototype frames, i.e., prototype frames that incorporate subjective estimations of the probability of attribute values. In section 5.1, I discuss the relationship between objective and subjective probabilities. The reason why we regard subjective probabilities to be a more suitable quantification of the typicality contribution of properties than productive frequency or property applicability comes from Schurz' (2001, 2005, 2007, 2011) work on the evolution-theoretic foundation of normality, which is summarised in section 5.2. In section 5.3, I summarise the arguments from Schurz (2012), where he shows how his account of normality can be applied to prototype concepts. The last section (5.4) demonstrates how we created an empirically testable model on this basis: probabilistic prototype frames.

5.1 Probability

Probability statements are an important component of everyday life and science. They are the basis for our decisions whenever outcomes are uncertain. We consider for example probability information like weather forecasts or accident statistics when we decide whether we should bring an umbrella or whether it is safe to take a plane.

Several interpretations of probability exist. Schurz (2015) distinguishes two main kinds: the objective or statistical probability of an event A, $\mathrm{pr}(\mathrm{A})$, which describes a property of reality, and the epistemic probability or grade of belief in A, $\Pr(\mathrm{A})$. Objective probabilities are derived from observed frequencies[18] and have the following properties: the probability of an event is a number between 0 and 1 inclusive, i.e., probabilities have an upper and lower limit. The probability of the negation of an event is the probability of the event subtracted from 1. If the probability that it will rain today is .8, then the probability that it will not rain is .2. The probability that an event or its negation occur is 1, it will certainly either rain or not rain, and the probability that an event and its negation occur simultaneously is 0, it is certain that it will not rain and not rain at the same time. These properties are summarised in the following axioms, also called Kolmogorov axioms, for events A and B (e.g., Schurz, 2015, p. 10):

- Non-negativity: $\mathrm{pr}(\mathrm{A}) \geq 0$.
- Normalisation: $\mathrm{pr}(\mathrm{A} \vee \neg \mathrm{A}) = 1$
- Finite additivity: $\mathrm{pr}(\mathrm{A} \vee \mathrm{B}) = \mathrm{pr}(\mathrm{A}) + \mathrm{pr}(\mathrm{B})$, if A and B are disjoint,

and from these theorems derived from them (Schurz, 2015, p. 11):

- Complementary probability: $\mathrm{pr}(\neg \mathrm{A}) = 1 - \mathrm{pr}(\mathrm{A})$
- Upper boundary: $\mathrm{pr}(\mathrm{A}) \leq 1$
- Contradiction: $\mathrm{pr}(\mathrm{A} \wedge \neg \mathrm{A}) = 0$.

A very important concept in this thesis is conditional probability, that is, the probability of an event A, given another event B occurs, formalised as $\mathrm{pr}(\mathrm{A}|\mathrm{B})$. For example, how probable it is

[18] For most practical applications, in particular for statistical laws, it would be more exact to speak of limiting frequencies. According to Schurz (2015, p. 60), statistical probabilities are most plausibly interpreted as generic propensities whose numerical value is determined by limiting frequencies that have inductive consequences for observable finite frequencies. An in-depth discussion is out of the scope of this thesis.

that something is sweet, when it is a fruit, $\mathrm{pr}(\mathrm{sweet}|\mathrm{fruit})$. Conditional probability is defined as the probability of A and B occurring together divided by the probability of B:

$$\mathrm{pr}(\mathrm{A}|\mathrm{B}) =_{\mathrm{def}} \frac{\mathrm{pr}(\mathrm{A} \wedge \mathrm{B})}{\mathrm{pr}(\mathrm{B})}, \text{if } \mathrm{pr}(\mathrm{B}) > 0.$$

Bayes' theorem holds for conditional probabilities (e.g. Schurz, 2015, pp. 14–15):

$$\mathrm{pr}(\mathrm{A}|\mathrm{B}) = \frac{\mathrm{pr}(\mathrm{B}|\mathrm{A}) \cdot \mathrm{pr}(\mathrm{A})}{\mathrm{pr}(\mathrm{B})}, \text{and}$$

$$\mathrm{pr}(\mathrm{A_i}|\mathrm{B}) = \frac{\mathrm{pr}(\mathrm{B}|\mathrm{A_i}) \cdot \mathrm{pr}(\mathrm{A_i})}{\sum_{1 \leq i \leq n} \mathrm{pr}(\mathrm{B}|\mathrm{A_i}) \cdot \mathrm{pr}(\mathrm{A_i})}.$$

Bayes' theorem is very useful for the calculation of unknown probabilities from those that are known. For example, if it is known how probable it is that sweet food items are fruit, $\mathrm{pr}(\mathrm{fruit}|\mathrm{sweet})$, and the prior probabilities of being sweet, $\mathrm{pr}(\mathrm{sweet})$, and being a fruit, $\mathrm{pr}(\mathrm{fruit})$, for food items[19] are known, Bayes' theorem allows the calculation of the unknown $\mathrm{pr}(\mathrm{sweet}|\mathrm{fruit})$. In its second formulation, it is important in scientific hypothesis testing in which B stands for evidence and $\mathrm{A_i}$ form a partition of hypotheses explaining the evidence. Furthermore, it is important for diagnostic reasoning, where B is taken to be an indicator for A. The higher $\mathrm{pr}(\mathrm{B}|\mathrm{A})$, the more sensitive is B as an indicator for A and the higher $\mathrm{pr}(\neg\mathrm{B}|\neg\mathrm{A})$, the more specific is B for A. The reliability of B, $\mathrm{pr}(\mathrm{A}|\mathrm{B})$, and its efficiency, $\mathrm{pr}(\neg\mathrm{A}|\neg\mathrm{B})$, can be calculated with Bayes' theorem from the known probabilities. (cf. Schurz, 2015, p. 15).

Objective probabilities are based on frequencies. Epistemic probabilities are understood as grades of belief. In formal epistemology, it is usually assumed that the Kolmogorov axioms are valid for them. The "principal principle", going back to Lewis (1980), states that the subjective probability of A, or rational grade of belief in A, should correspond to the objective probability of A – if you know that the objective probability of A is r, then your subjective estimation should be r as well, if you are a rational agent:

$$\mathrm{Pr}(\mathrm{A}|\mathrm{pr}(\mathrm{A}) = \mathrm{r}) = \mathrm{r}.$$

Schurz (2015, p. 73) calls it the singular coordination principle. He extends it to the statistical coordination principle (StC) as follows, where Gx corresponds to the sentence "x is/has G" and $\mathrm{E}(\mathrm{b_j})$ is the knowledge about other individuals $\mathrm{b_j}$:

(StC) 1) For a statistical hypothesis H that probabilistically implies $\mathrm{pr}(\mathrm{Gx}) = \mathrm{r}$:

$$\mathrm{Pr}\big(\mathrm{Ga_i}\big|\mathrm{H} \wedge \mathrm{E}(\mathrm{b_1}, \ldots, \mathrm{b_n})\big) = \mathrm{r}, \text{if } \mathrm{a_i} \neq \mathrm{b_j} \text{ for all } \mathrm{j} \in \{1, \ldots, \mathrm{n}\}.$$

2) For a statistical hypothesis H that probabilistically implies $\mathrm{pr}(\mathrm{Gx}|\mathrm{Fx}) = \mathrm{r}$:

$$\mathrm{Pr}\big(\mathrm{Ga_i}\big|\mathrm{H} \wedge \mathrm{Fa_i} \wedge \mathrm{E}(\mathrm{b_1}, \ldots, \mathrm{b_n})\big) = \mathrm{r}.$$ (Schurz, 2015, pp. 74–75)

[19] Food items are the narrowest reference class in this example. According to Reichenbach's (1949) principle of the narrowest reference class, the subjective probability of an event is determined as its estimated probability to occur in the narrowest reference class from which the subject knows that A belongs to it. (Schurz (2015, p. 7), pp.92-93)

The StC states that the epistemic probability of a_i having G corresponds to the objective probability r if this objective probability is known.

In my thesis, I focus on conceptual representations of real agents, contrary to the ideal, rational agents which are usually assumed in formal epistemology. In what follows, I will call the subjective estimations of probabilities made by real agents subjective probability. Contrary to their rational counterparts, real agents do not necessarily follow the laws of probability or rationality conditions from formal epistemology. Instead, they tend to both over- and underestimate probabilities, which was famously shown in a series of experiments by Amos Tversky and Daniel Kahneman, summarised in Kahneman et al. (1982).

Kahneman and Tversky (1972) report reliable, systematic deviations from the probability axioms. Events that represent the parent population and its salient properties were rated to be more likely than probability theory would predict. For example, the birth order BGBBBB (boy, girl, ….) was rated to be less likely than GBGBBG. They explain this by the fact that in the population, girls and boys are more equally distributed than in the first order, which makes the second order more representative of the parent population. (p. 432).

In Tversky and Kahneman (1982), possible explanations for the systematic deviation are presented. The first reason is representativeness: when asked if A is B, they report that participants tend to ignore prior probabilities of outcomes, as well as the sample size, predictability and chance and are instead guided by the representativeness of A for B – if A is representative for B, it is categorised as B. For example, they explain that the coin toss sequence H-T-H-T-T-H (head, tail, …) is rated to be more probable than the sequence H-H-H-T-T-T because the former is intuitively more representative of randomness. Another famous example involves reasoning about the profession of Linda, described as:

> "31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations." (p. 92)

Linda was rated to be more likely a feminist bank teller than just a bank teller, which they explain by her being more similar to a feminist than to a bank teller. As all feminist bank tellers are bank tellers, this is implausible in view of the probability axioms, because it states $\Pr(A \wedge B) > \Pr(A)$.

The second reason for the deviation from objective probability laws they discuss is availability, which explains cases in which the probability of A was rated to be higher than B when it was easier to find examples for A. For example, as it is easier to find examples for words that begin with an r than for words that have r as their third letter, people judge words that begin with an r to be more frequent than those with r as their third letter. The third reason is adjustment and anchoring for tasks in which participants' judgments varied depending on which initial value was presented to them. When asked to compute $8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$ in 5 seconds, the median estimate was 512 and for $1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8$ it was 2,250[20] (Tversky & Kahneman, 1982, p. 15).

Keeping the biases and heuristics identified by Tversky and Kahneman in mind is certainly important when interpreting subjective probabilities. Schurz (2007) argues that the results they report do not show that humans do not reason well with probabilities at all, but that there are

[20] The correct answer is 40,320.

only certain reliable rules for probabilistic reasoning and they "perform badly in other probability tasks which are not covered by these rules" (Schurz, 2007/2007, p. 629). The reliable rules are summarised in section 5.2.

There is a second explanation why objective and subjective probabilities can differ. Figure 31 illustrates that, as people only observe parts of the environment, this can lead to a more or less accurate sampling of the frequencies and because of this to different probability estimations in different individuals. The first person, S_1, has observed the property A in 3 individuals of category C, c, g and h, and has observed only one member of C that did not have A, i. The probability that a member of C has property A according to this evidence is .75. For S_2, it is .5 because they observed the property A in half of the members. The objective probability for the whole population is .56 which means that S_2 has observed a more representative sample. As mentioned in the introduction, it is not necessary for theories of concepts to reflect the external world and thus objective probabilities, but to explain the intersubjective stability and ability of successful communication it is necessary that the estimations of different individuals are similar. That this should be the case in the domain of prototype theory is argued for in the sections 5.2 and 5.3.

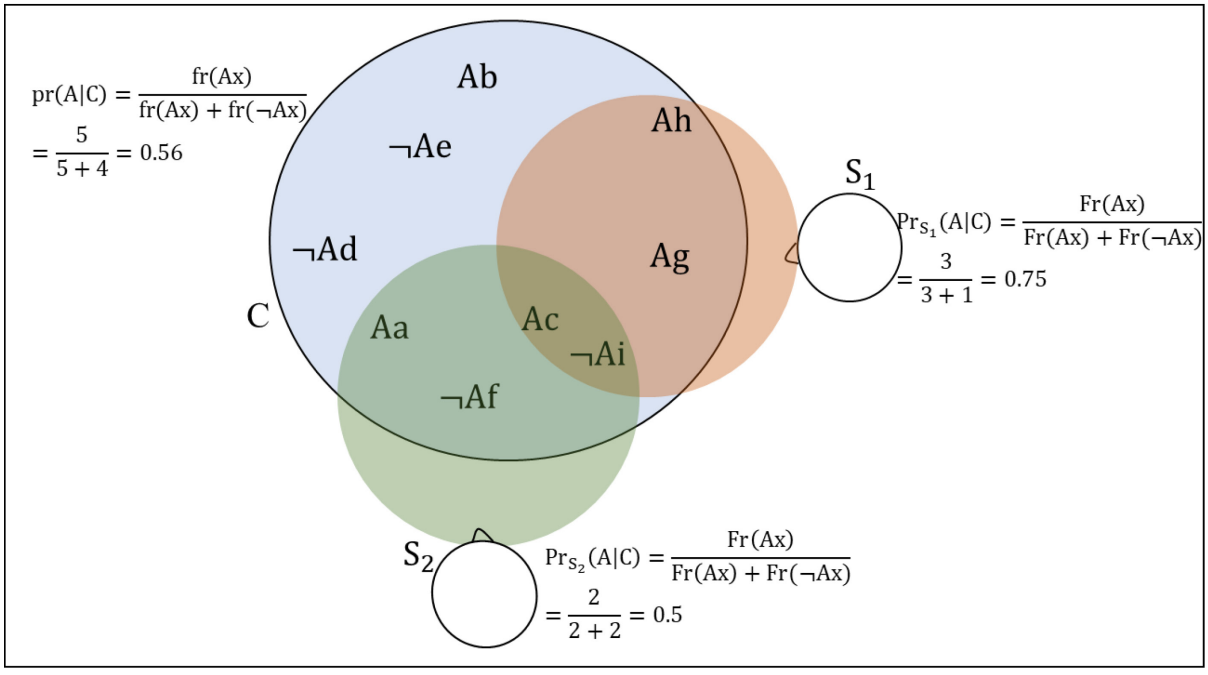

Figure 31: Objective and subjective probability of the property A in a category C from two subjective viewpoints.

In my experiments, I collected subjective probability estimations for properties of categories for which the samples different people observed should be similar. The question whether these estimations correspond to the objective probabilities is difficult to answer, because objective probabilities are difficult to know in the investigated domain:

> "The numerical probabilities associated with prototypical properties are typically highly context-sensitive, so that we are unable to specify them without specifying the context. For example, what percentage of all birds can fly (when?, where?)" (Schurz, 2012/2012, p. 531).

A high intersubjective stability of the ratings would support the assumption that the ratings are based on similar observations and correspond roughly to the (unknown) objective probabilities. The reverse is not true: high intersubjective variability could indicate that participants interpreted the question in different ways, that they observed different environments in which

the distributions of properties differ or that they estimate probabilities in a roughly uniform environment differently. I discuss the general rating behaviour of participants in section 6.4.1, their relationship to the laws of probability in section 6.4.2 and an application of Bayes' theorem to subjective probabilities in section 6.4.3.

5.2 Schurz' evolutionary account of normality

Normic laws are normality statements like "As are normally Bs", formalised as A ⇒ B. They are different from strict laws in that they are not falsifiable by one counterexample, because they imply the existence of exceptions to the law – that birds can normally fly implies that there are birds that cannot fly. Observing a bird that cannot fly would falsify the strict law "All birds can fly" but not the normic law "Birds can normally fly". Given the ubiquity of normic laws in everyday life and several scientific disciplines, like biology and the humanities, it is desirable to assign them falsifiability conditions and to specify their meaning. This is what Gerhard Schurz (2001) does in his article "What is 'normal'? An evolution-theoretic foundation for normic laws and their relation to statistical normality" and subsequent work. To explain the "omnipresence, lawlikeness, and reliability" (p. 476) of normic laws, they have to be interpreted as implying statistical majority, he argues. That birds can normally fly then implies that most birds can fly. Statistical laws are statements like "r% of As are Bs" and they can be weakened and ultimately falsified by many observations that lie outside of their acceptance interval (usually the 95% confidence interval around r). Normic laws do not specify a number for r, only that it is high, at least higher than .5. They can be falsified by observations that show that r is not the statistical majority. Normic laws like "birds can normally fly" resemble generics like "birds can fly". There is interesting research on generics and their interpretation. In some cases, like "birds lay eggs" or "mosquitoes carry the West Nile virus", statistical majority is not necessary for people to judge a generic to be true (e.g., Leslie, 2008).

The connection between normality and statistical majority is stated in the statistical consequence thesis:

> (StC) A ⟹ B implies that the conditional statistical probability of B given A, pr(B|A), is high. (Schurz, 2005, p. 38).

The ontological justification of StC is that normic laws are not accidental: the statistical majority is a result of the evolutionary history of the entities they describe. Normic laws imply statistical majority whenever they describe evolutionary systems. In the generalised theory of evolution, the term evolutionary system refers to the entities that undergo evolution. They are organisms in the case of biological evolution and human communities in cultural evolution. The basic terms of generalised evolution theory and their counterparts in biological and cultural evolution are in Table 20. It illustrates that both kinds of evolution share the same general characteristics.

Table 20: Basic terms of the generalised theory of evolution and their counterparts in biological and cultural evolution (Schurz (2021)).

generalised evolution	biological evolution (BE)	cultural evolution (CE)
evolutionary systems	organisms in their environment	humans and their societies CE takes places on top of BE
reprones	genes in the cell-nucleus, epigenetic patterns	memes / acquired information, software of the brain
phenetic traits	organs, abilities, niche construction	behavioural & cognitive skills, technologies, language, ideas and knowledge
reproduction	replication, DNA copy	imitation of learning between humans semantic reproduction
inheritance system	sexual (diploid)	asexual (blending inheritance)
variation	mutation and recombination	interpretation and variation of transmitted memes
selection	higher rates of reproduction (due to different causes)	
	of genes (and epigenetics)	of memes

Evolutionary systems self-regulate, this means that most of their individuals are in a prototypical norm state for the majority of the time that ensures their survival[21]. Normic laws about evolutionary systems imply statistical majority because they describe this prototypical norm state: they are "the phenomenological laws of self-regulatory systems" (Schurz, 2001, p. 479). Self-regulation is possible because of three characteristics that are common to all evolutionary processes:

(E1) a mechanism of *variation* which acts in larger populations of evolutionary systems which are in mutual competition.

(E2) a mechanism of *reproduction* which leads to consecutive generations of evolutionary systems – hence variations must be heritable, and

(E3) an environment which *selects* the fittest among the variations, i.e., those with the highest reproduction rate – hence variations must differ in their fitness. (Schurz, 2001, p. 481)

Variation is the reason why normic laws allow for exceptions: a part of evolution is the production of individuals with alternative properties, through mutation and recombination in biological evolution and via interpretation and variation in cultural evolution.

[21] In biological evolution, this literally means surviving. Schurz clarifies that for example in the cultural evolution of electronic devices, this could mean "surviving on the economic market" (Schurz (2001, p. 480)).

While normic laws are not physically necessary, because they result from accidental circumstances, they are a necessary result of the evolutionary process (p.482). Not all prototypical properties bring a direct selective advantage. Some are side-effects of those that bring one. Schurz calls the former fundamental prototypical properties and the latter derived prototypical properties. For example, it is a fundamental prototypical property of hearts to circulate blood, but a derived one to make the characteristic sound of a heartbeat. Both fundamental and derived properties are the results of a common cause, lying in the genotype of living organisms and the memotype of cultural artifacts (p. 485). This is summarised in his definition of prototypicality:

> (PN) For S a class of evolutionary systems, and T a trait of S-members:
>
> (i) T is a prototypical trait of S-members at time t iff T is produced by a reprotype R and from T's first appearance in the S-history until time t, there was overwhelming selection in favor of R.
>
> (ii) T is a fundamental prototypical trait of S-members at time t iff the selection mentioned in (i) was overwhelmingly caused by R's producing T. T is a derived prototypical trait of S-members at time t iff the selection mentioned in (i) was overwhelmingly not caused by R's producing T. (Schurz, 2001, p. 494)

Together with the assumption that "T is a fundamental prototypical trait of S-members *simpliciter* iff T is a fundamental prototypical trait of S-members at the *latest* time-point of S's existence" (p. 494), the following conclusions can be drawn:

> Conclusion 1: If T is a prototypical trait of S-members at time t, then there exists a reprotype R which produces T, and from T's first appearance until t there was overwhelming selection in favor of R. [From definition PN.]
>
> Conclusion 2: If T is a prototypical trait of S-members at t and t is not a time soon after T's first appearance, *then* for most time points from T's first appearance until t, (i) T was a prototypical trait of S-members and (ii) most S-members possessed reprotype R and, therefore, trait T. [(i) from conclusion 1 and definition PN; (ii) from conclusion 1, definitions of "production", "selection", and probability theory.] (Schurz, 2001, p. 494)

These conclusions lead together with the premise that "Most classes of evolutionary systems do not become extinct soon after acquiring a selectively advantageous trait" (p.495) to the following version of the statistical consequence thesis:

> (StC) For most classes of evolutionary systems S and times t of their existence it holds that if T is a (fundamental or derived) prototypical trait of S-members at time t, then most S-members will possess trait T at time t. (Schurz, 2001, p. 495).

StC ensures that the Default Modus Ponens is a reliable inference for normic laws:

(MP): $Ax \Rightarrow Bx, Aa \mid\sim Ba$, provided the premise set contains all information that is known and probabilistically relevant for the conclusion Ba.

where $\mid\sim$ is the non-monotonic inference relation and the premise set has to satisfy the total evidence condition (Schurz 2005, pp. 40–42). For example, from "birds normally fly" and "a is a bird" it can be inferred that a can probably fly, as long as nothing else of relevance is known about this bird that could constitute an exception. The system P introduced in Adams (1975) has more rules than MP. They are however only reliable for probabilities that are only infinitesimally smaller than 1, which does not apply for the normic conditionals found in practical life (Schurz, 2005, p. 44). With the additional assumption that the conclusions drawn from normic conditionals have only few premises, the probability of the conclusion will be high for inferences derived by the following rules (excerpt, Schurz, 2005, p. 42, Schurz, 2015, p. 34):

Cautious Transitivity(CT): $A \Rightarrow B, A \wedge B \Rightarrow C \mid\sim A \Rightarrow C$
Cautious Monotonicity (CA): $A \Rightarrow B, A \Rightarrow C \mid\sim A \wedge B \Rightarrow C$
Weak Rational Monotonicity (WRM): $A \Rightarrow B, \neg(A \Rightarrow \neg C) \mid\sim A \wedge C \Rightarrow B$

The inference WRM is needed because monotonicity is obviously not valid in non-monotonic logics – the introduction of counterexamples is explicitly possible. It is an explanation why default inheritance is reliable, whenever the normic conditional has a high probability and the SC in question is not known to be a counterexample. The probabilistic inequality with uncertainty U defined as 1 minus probability, $U(B|A \wedge C) \leq \frac{U(B|C)}{Pr(A|C)}$, applies which means that as long as $Pr(A|C)$ is high and $U(B|A)$ is low, the conditional uncertainty of the conclusion will be low (Schurz, 2012/2012, p. 552). For example, if "birds can normally fly" (first premise of WRM) and I know that "it is not the case that birds are normally not small" (second premise of WRM), I can reliably conclude that "small birds normally fly" (conclusion of WRM).

In Schurz (2005), he reports that MP was used by participants and that the addition of "normally" or "mostly" to the premises only led to slightly more cautious replies, but in the predicted direction (p. 48). Furthermore, the law of specificity, which involves reasoning with an instance known to be exceptional, for example in

> Chocolate tastes sweet.
> By mistake this chocolate cake was baked with salt instead of sugar.
> Therefore: This chocolate cake tastes sweet. (Schurz, 2005, p. 49)

was found to be respected by the participants[22] and rated differently than clearly contradictory premises, which it would not be if people applied classical reasoning (pp. 48-49). In Schurz (2007), similar results are reported. Given that participants treated conditionals as "high conditional probability assertations" (ibid., 629) and in view of other studies that report similar results (Oberauer & Wilhelm, 2003, Evans, Jonathan St. B. T. et al., 2003, Pfeifer & Kleiter, 2005), he concludes that humans generally follow the non-monotonic inference patterns for normic laws, but only for descriptive conditionals. For normative conditionals like "drinking alcohol is only allowed for adults", people interpret conditionals as strict and thus follow the strict pattern Modus Tollens from classical logic (Cosmides & Tooby, 1992). The restriction of

[22] Mean = 3.03, SD = 1.41 on a scale from 4 (yes) to 0 (no).

probabilistic reasoning to domains in which they were evolutionary selected is also found in Gerd Gigerenzer's account of ecological rationality (e.g. Gigerenzer, 2000).

In summary, the structure of the world was shaped by evolution in a way that leads most entities we encounter to not follow strict, but normic laws. Normic laws have statistical content whenever they describe entities that were shaped by evolution and are therefore falsifiable. Humans are adapted to reasoning with normic laws in a way that follows the laws of non-monotonic reasoning.

5.3 Evolutionary normality and concepts

Schurz (2012) argues for conceptual pluralism according to which the best way to specify the meaning of a word depends on its domain. Some meanings could be stated as a list of singly necessary and jointly sufficient conditions, some with a background theory and some with prototype theory. He identifies the appropriate domain for prototype concepts as the domain of evolutionary systems, because in this domain prototypical normality is connected to statistical majority. As our living environment is filled with evolutionary systems, "human cognition is well adapted to prototypicality structures" (p. 531).

Schurz notes that typicality statements like "Cs have typically P" are a special kind of normic laws and they have objective meaning if StC is assumed. Applied to conceptual representations and assuming that subjective probability is similar to objective probability, the statistical consequence thesis states that the subjective probability, Pr, of P given C, (Pr(P|C)), is high for typical properties of C, if the majority of SCs of C is observed to have P.

To the popular objections against prototype theory, he replies that prototype theory is a valuable theory of cognition, when its scope and the domain of its application are restricted. He makes this clear in two theses. The first identifies the scope of prototype theory as providing a means for predictive and diagnostic reasoning with common-sense concepts (p. 542):

> (P1) One (if not *the*) major evolutionary function of cognition is efficient *predictive* reasoning (inferring the effects of practically important causes) and *diagnostic* reasoning (inferring the causes of practically important effects). *Categorisation* is a *necessary condition* of predictive and diagnostic reasoning, but categorisation *per se* is not evolutionarily advantageous because not every categorisation is predictively and diagnostically efficient.

In agreement with Rosch (1978), he identifies predictively and diagnostically efficient categorisations as those that "possess computationally simple categories which figure as junctions in a dense system of lawlike connections" (ibid.).

In his second thesis, he specifies the proper domain of prototype theory and differentiates two kinds of prototypicality of properties for a species S, namely in the wide sense (i.w.s.) and in the narrow sense (i.n.s.)[23] (p. 534):

[23] In what follows, properties that are prototypical i.n.s. in Schurz' terminology are referred to as diagnostic properties, and properties that are prototypical i.w.s. are called frequent.

(P2) The proper domain of prototype theory is evolutionary systems […]. In this domain, predictive and diagnostic efficiency is achieved because […] evolutionary systems obey the following principles:

(P2.1) Each species (or kind) S of an evolutionary system is characterized by a bundle of prototypical properties i.w.s. which have been selected during the evolution of (ancestors of) S and which S-members possess with a high statistical frequency.

(P2.2) A certain subset of these prototypical properties i.w.s. – namely the prototypical properties i.n.s. – are highly discriminative vis-a-vis sibling species of S, because each kind of evolutionary system had its specific adaptation history to the selection requirements of its environment.

(P2.3) All kinds of evolutionary systems include *exceptional* exemplars or subkinds, which deviate from the prototype pattern of the kind.

(P2) states that for any prototypical property (i.w.s.), $\mathrm{pr}(P|C)$, is high and for prototypical properties i.n.s., i.e., diagnostic properties, additionally $\mathrm{pr}(C|P)$ is high. Schurz notes that each of m diagnostic properties justifies the inference that the object in question belongs to C, $P_i \Rightarrow C$, and from the presence of C, all of the n (n>m) prototypical properties can be inferred, $C \Rightarrow P_j$, for $j \neq i$. C can be seen as "a mediator in a network of $m - (n - 1)$ direct predictive or diagnostic inferences $P_i \Rightarrow P_j$" (p. 544).

He identifies Cs for which the sum of diagnostic properties is high as basic level categories and gives an evolution-theoretic explanation for them: "their branch of ancestors in the tree of evolutionary descendance has a *long, homogeneous,* and *category-specific* selection history which produced many prototypical properties i.n.s." (p. 544).

With the evolutionary foundation, the intersubjective meaning stability of prototype concepts can be explained by the world being structured in a way that leads to similar observations from different individuals. The fact that category membership cannot be deduced from typicality for a category is outside of the scope of prototype theory, which allows for the existence of exceptions, as long as there is a statistical majority of cases in the prototypical norm state. Schurz proposes to introduce an analytic core meaning that decides about membership.[24] Despite being based on world-knowledge, he considers prototypes as part of word meaning: "If one wants a psychologically realistic notion of meaning, which reflects that content which natural language speakers immediately associate when parsing the utterance of linguistic expressions (cf. Springer and Murhpy, 1992), then prototypes *should* be regarded as *part* of the meaning." (p. 546). (cf. pp.545-546).

5.4 Frame adaptation

With the terms introduced in the foregoing chapters, a more specific version of the frame hypothesis FH can be given:

[24] Another option is to assign weights to typical properties and to define a membership threshold, e.g. Hampton (1995)

(PFH) Quantified prototype frames are the format of mental representations for concepts referring to categories in the prototype domain, that is, for evolutionary systems.

In personal communication with Gerhard Schurz (2016-2019), Schurz' (2012) proposal was used as a basis to construct empirically testable prototype frames. A short summary of the formulae presented here was published in Schuster et al. (2020).

The PFH (short for probabilistic frame hypothesis) can be divided into five hypotheses. The first hypothesis summarises the connection between typicality and prototypical properties from Schurz (2012).

1) The typicality of SCs for C is determined by the probability that the SCs have prototypical properties P of C.

The second hypothesis specifies the two conditional probabilities that are high for properties that contribute to the prototypical norm state of evolutionary systems.

2) Prototypical properties are those properties brought about in C's evolutionary history that
   a) are possessed by SCs with high statistical frequency, $\mathrm{pr}(\mathrm{P}|\mathrm{C}) = \mathrm{high}$, allowing for the inference $\mathrm{C} \Rightarrow \mathrm{P}$, termed frequent or prototypical in the wide sense (iws),
   b) are additionally discriminative with regard to contrast categories of C, $\mathrm{pr}(\mathrm{C}|\mathrm{P}) = \mathrm{high}$, allowing for the inference $\mathrm{P} \Rightarrow \mathrm{C}$, termed diagnostic or prototypical in the narrow sense (ins).

The third hypothesis specifies the relationship between objective and subjective probabilities: the objective probabilities are estimated, and these estimations are integrated into the conceptual structure. It makes a connection between cognition and world in assuming that the property probability distributions found in evolutionary systems are represented in the mind and guide typicality ratings:

3) Humans build up a conceptual structure which incorporates subjective estimations that are based on the observation of these probabilities, formalised as $\Pr(\mathrm{P}|\mathrm{C})$ and $\Pr(\mathrm{C}|\mathrm{P})$.

The fourth hypothesis specifies that frames are an adequate tool to represent conceptual structure and specifies how the probabilities are embedded in frames: as weights on different levels of the frame. In accordance with the frame hypothesis, properties are represented as structured into n attributes $\mathrm{A_i}$ with $\mathrm{m_i}$ values $\mathrm{V_{ij}}$. Each attribute value's contribution to category typicality is quantified with a subjective estimation of the value's frequency and diagnosticity, defined, as introduced above, as $\Pr(\mathrm{V}|\mathrm{C})$ and $\Pr(\mathrm{C}|\mathrm{V})$ respectively.

The question "How typical is SC for C?" is, as commonly accepted, interpreted as the question "How similar is SC to the prototype of C?". As a measure of the similarity between SC and C in regard to a given attribute with $\mathrm{m_i}$ values, we chose the sum of the minimum value probabilities of SC and C:

$$\mathrm{Sim}(\mathrm{SC}, \mathrm{C}|\mathrm{A_i}) = \sum_{\mathrm{j}=1}^{\mathrm{m_i}} \min\left(\Pr\left(\mathrm{V_{ij}}\middle|\mathrm{C}\right), \Pr\left(\mathrm{V_{ij}}\middle|\mathrm{SC}\right)\right).$$

The explanation for this measure is as follows: if the probabilities of the values of C and SC would be identical (i.e., ideally similar), then their min would correspond to the probabilities and the sum of the mins would be 1, i.e., maximal. However, if these values differ largely, then the min of the two would for each value be much smaller than the max of the two, and therefore, the sum of their mins would be small, much smaller than 1, in the least similar case even 0. Therefore, the sum of the mins is an elegant measure for the similarity.

Value diagnosticity reflects prototypicality ins and is the probability that an object belongs to the C in question, given it has the attribute value in question, $\Pr(C|V_{ij})$. Attribute diagnosticity is calculated from value diagnosticity as the maximum of the value diagnosticities for each attribute. This is based on the assumption that attributes which contain a highly discriminative value should be more salient and receive a higher weight than attributes that have a flat probability distribution over all values (i.e., in which each value is roughly equally probable). All these thoughts are summarised in the fourth hypothesis:

4) The structure of prototype concepts is best described by frames. A frame has several attributes $A_1, A_2, \dots, A_n$. Each attribute A_i has several possible values $V_{i1}, V_{i2}, \dots, V_{im_i}$. Properties correspond to values, property dimensions to attributes. The attributes and values receive weights that reflect their typicality contribution in the following way:

   a) probability: the probability of a value conditional on a (sub)category reflects the prototypicality i.w.s. of that value for the (sub)category, estimated for each value V_{ij}. The attribute similarity of a SC to C is the sum of the minimum value probabilities for each attribute value:

   $$\mathrm{Sim}(SC, C|A_i) = \sum_{j=1}^{m_i} \min\left(\Pr(V_{ij}|C), \Pr(V_{ij}|SC)\right)$$

   b) diagnosticity: the diagnosticity of an attribute A_i reflects the prototypicality i.n.s. of its values and is defined as the maximum of the inverse value probabilities, $\Pr(C|V_{ij})$, and normalised by the sum of these maxima

   $$\mathrm{diag}_{A_i}(C) = \frac{\max\left(\Pr(C|V_{i1}), \dots, \Pr(C|V_{im_i})\right)}{\sum_{i=1}^{n} \max\left(\Pr(C|V_{i1}), \dots, \Pr(C|V_{im_i})\right)}.$$

The fifth and final hypothesis presents the formula to calculate typicality. Typicality is defined as the weighted average of the similarities per attribute:

5) The typicality of a subcategory SC for a category C is defined as the weighted average of their similarities for each attribute (according to 4a) above); weighted by the attribute diagnosticities (according to 4b) according to the formula:

$$\mathrm{typ}(SC, C) = \sum_{i=1}^{n} \mathrm{diag}_{A_i}(C) \cdot \mathrm{Sim}(SC, C|A_i)$$

The maximal similarity per attribute is always 1, which is the case for any SC that corresponds to C in all probabilities. The sum of similarities per attribute is weighted with the diagnosticity

of this attribute. The normalisation of diagnosticity ensures that the maximal typicality is 1 and is reached if the SC has equal or higher subjective probability ratings for all attribute values in question.

The probabilistic prototype model assumes that categories are mentally represented in a way that is describable by prototype frames quantified with subjective probability estimations for all attribute values in terms of their diagnosticity and probability. Typicality ratings for SCs are then explained as a comparison per attribute of the SC frame, quantified with the probability of each attribute value, with the C frame. The results of each comparison are weighted with the attribute diagnosticity. The sum of all attribute-weighted similarities predicts the typicality of the SC.

Table 21 shows an example for typicality calculations of *apple* and *avocado* for *fruit* with two attributes, COLOUR and TASTE, the former with 5 and the latter with 3 values, all quantified with fictional probabilities. On the left side are subjective probabilities of each value both for C and the SCs. On the right side are the reversed probabilities for *fruit*, on the basis of which diagnosticity is calculated as the maximum per attribute normalised by dividing by all maxima. For example, the diagnosticity of TASTE is calculated as:

$$\mathrm{diag}_{\mathrm{A_{TASTE}}}(\mathrm{fruit}) = \frac{\max(0.9,0.1,0)}{\max(0.9,0.1,0) + \max(0.2,0.2,0.2,0,0)} = \frac{0.9}{0.9 + 0.2} = 0.82.$$

Then, the similarities per attribute are computed as the sum of the minima of the values, shown below the probabilities in Table 21. The last step is to calculate typicality as the diagnosticity-weighted average. In these example calculations, apple has a high typicality for fruit of .9 and avocado has a very low typicality of .04, which reflects that the value probability distributions for fruit and avocado are very different with regard to colour and taste.

Table 21: Example for fruit typicality calculations with probabilistic prototype frames with fictional values.

		Pr(P|C) and Pr(P|SC)			Pr(C|P)		
attribute	value	fruit	apple	avocado	fruit	max	$\mathrm{diag}_{\mathrm{A_i}}(\mathrm{C})$
TASTE	Sweet	0.75	0.8	0	0.9	0.9	0.82
	Tart	0.24	0.2	0	0.1		
	Other	0.01	0	1	0		
COLOUR	Red	0.35	0.6	0	0.2	0.2	0.18
	Yellow	0.3	0.2	0	0.2		
	Orange	0.2	0	0	0.2		
	Green	0.1	0.2	0.5	0		
	Other	0.05	0	0.5	0		

$$\mathrm{Sim}(\mathrm{apple}, \mathrm{fruit}|\mathrm{TASTE}) = \min(0.75,0.8) + \min(0.24,0.2) + \min(0.01,0) = 0.9$$
$$\mathrm{Sim}(\mathrm{apple}, \mathrm{fruit}|\mathrm{COLOUR}) = 0.65$$
$$\mathrm{typ}(\mathrm{apple}, \mathrm{fruit}) = 0.82 \cdot 0.9 + 0.18 \cdot 0.65 = 0.9$$

$$\mathrm{Sim}(\mathrm{avocado}, \mathrm{fruit}|\mathrm{TASTE}) = \min(0.75,0) + \min(0.24,0) + \min(0.01,1) = 0.01$$
$$\mathrm{Sim}(\mathrm{avocado}, \mathrm{fruit}|\mathrm{COLOUR}) = 0.15$$
$$\mathrm{typ}(\mathrm{avocado}, \mathrm{fruit}) = 0.82 \cdot 0.01 + 0.18 \cdot 0.15 = 0.04$$

An interesting variable in these calculations is each values' typicality contribution, typcon, which is the minimum of the value in question multiplied with its attribute's diagnosticity:

$$\mathrm{typcon}(V_{ij}|SC) = \mathrm{diag}(A_i) \cdot \min(\Pr(V_{ij}|SC), \Pr(V_{ij}|C))$$

For example, the typicality contribution of the value Sweet for TASTE is $\mathrm{typcon}(V_{\mathrm{TASTE,Sweet}}|\mathrm{apple}) = 0.82 \cdot 0.75 = 0.61$, which corresponds to the maximal typcon for fruit, computed by using the category probabilities instead of the minimum. Figure 32 shows all typcons for the example calculations. This way of presentation makes it easy to determine in how far the different value probabilities contribute to typicality and how similar the SC probability distributions are to the C distribution.

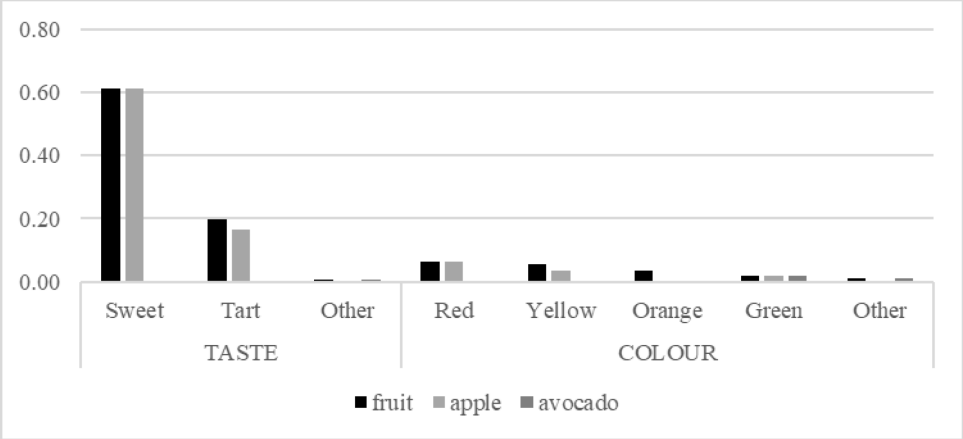

Figure 32: Typicality contributions (typcon) for apple and avocado and maximal typicality contributions for fruit for the example calculations

Similar ideas are found in Corter and Gluck (1992), who also use a probabilistic measure to represent categories and properties. They quantify the category utility CU, which is high when a category allows us to predict and communicate information about the properties of SCs. Their measure employs a set of properties P, their probabilities in the category, Pr(P|C), which they term category validity, and Pr(P) and Pr(C), the overall base rate of property P:

$$\mathrm{CU}(C, P) = P(C) \sum_{k=1}^{m} [\Pr(P_k|C)^2 - \Pr(P_k)^2]$$

While their question differs from ours, the variables they consider important are very similar to what we propose here.

6 Experiments

To test the cognitive plausibility of the probabilistic prototype frames developed in the foregoing chapter, subjective estimations of the property probabilities used in the formula were gathered in 3 experiments: the probability of attribute values (the probabilities with which both SCs and Cs have prototypical properties), and the diagnosticity of attribute values (the probability that something is in C given it has these prototypical properties).

Originally, it was planned to collect all these estimates in one, very large survey. However, after an advisory meeting with James Hampton, this plan was altered to avoid the collection of probability judgements that are not contributing to typicality according to the formula, as this would tire the participants without gain. The experiment was split into 3 parts that could be completed in 10 to 40 minutes each. Based on the assumption that all prototypical properties should be diagnostic for the category, because even a high probability multiplied with a low diagnosticity leads to a low typicality contribution, diagnosticity ratings were collected first, with the goal to reduce the number of properties for which probability judgements for SCs are needed. For a noteworthy typicality contribution in the model, it is additionally necessary that the properties are frequently found in the category in question. Thus, in the second experiment I collected category probability ratings for the identified diagnostic properties. In the final and largest experiment, I then collected SC property probabilities for the small number of properties that were found to be both diagnostic and frequent for the investigated Cs.

I chose the concepts *fruit* and *vegetable* for this investigation for two reasons. The first reason is formal: they are clear contrast categories, meaning that in the narrowest reference class, each SC is either a fruit or a vegetable[25] and category-membership in them is mutually exclusive – something cannot be a fruit and a vegetable at the same time. These properties are useful to determine whether participants' ratings follow the formal laws of probability, like Bayes' theorem, because the probabilities conditional on ¬C can be replaced with vegetable probabilities for fruit and with fruit probabilities for vegetable. The second reason is empirical: the categories were investigated in almost all typicality studies which means that there is a broad experimental background for comparison (see section 3.2). Additionally, Smith et al. (1988) made a very similar investigation to the one described here with these categories (see section 8.2).

This chapter presents successively the first (6.1), second (6.2) and third (6.3) experiment. Each chapter has sections for participants, stimuli, design and procedure, results and discussion. A general discussion is in section 6.4. It discusses how the participants behave in subjective probability estimation tasks in general, how the results from the different experiments are related, and presents results on applying Bayes' theorem on the results of experiment 1 and 2.

6.1 Experiment 1

The goal of experiment 1 was to gather diagnosticity ratings for category properties (or values) of *fruit* and *vegetables* and identify the most diagnostic properties for each category. As presented in chapter 5, a property's diagnosticity for a category is understood as the conditional

[25] Though Hampton (1988a) found that the composed concept *fruit-or-vegetable* included SCs that were rated to be neither a *fruit* nor a *vegetable*. There are borderline cases (e.g., olives and tomatoes) that, according to empirical data, are in both categories. The prototypes are however clearly contrasting and there is no third category in the narrowest reference class.

probability Pr (C|P), i.e., the probability that something belongs to that category, when the property in question is known to apply to it. Diagnosticity is closely related to salience – properties that aid categorisation can be assumed to be more prominent in the mental representation of a concept.

The participants are described in 6.1.1, the way how stimuli were chosen in 6.1.2, the design and procedure in 6.1.3, the reliability of the results in 6.1.4 and their descriptive statistics in 6.1.5. Section 6.1.6 discusses the results.

6.1.1 **Participants**

The general information on the participants is summarised in Table 83 in the appendix. Participants were 30 paid volunteers recruited via the platform Prolific (19 female, 11 male, mean age 30.8, SD 7.8). 20 live in the UK, 7 in the USA and 3 in Canada. 18 of them characterised their dialect as British, 10 as American and 2 as other. 29 said that they have no knowledge of botany. All except one subject have at least high school education, most did some college (7) or a bachelor (8).

6.1.2 **Stimuli**

A key assumption of prototype theory that is made explicit in the family resemblance hypothesis (FRH, section 4.1) is that the prototypical properties of Cs are those that are shared between many SCs. The identification of relevant properties for the experiments was therefore based on property lists for SCs which are already available in the literature.

Properties were taken from property generation data from McRae et al. (2005) and Devereux et al. (2014). McRae et al. (2005) provide property lists for 541 English nouns chosen to cover words that were used in previous experiments. 30 participants from the McGill University, the University of Southern California or the University of Western Ontario listed properties for each concept. They were provided 10 blank lines and asked to provide different types of properties. The data available online[26] contains all properties with a PF of at least 5.

Devereux et al. (2014) provide property lists for 866 concrete concepts, including all from the McRae et al. data which are also common in British English (490), and additional ones from previous experiments and to increase the distinctiveness of the norms. They had 123 participants from the Centre for Speech, Language and Brain subject pool (University of Cambridge), and 30 participants generated properties for each concept anonymously online. The data they provide online[27] includes all properties that were generated by at least 2 participants.

For an identification of fruit and vegetable properties, the lists were filtered to include only words for which "a fruit"/"is a fruit" or "a vegetable"/"is a vegetable" were generated by at least 4 participants[28]. In the McRae et al. (2005) dataset, 29 fruit and 27 vegetable SCs were identified. Avocado (17F, 8V), olive (9F/5V), pumpkin (8F/8V), rhubarb (8F/16V) and tomato (21F/11V) are borderline cases that are included in both categories. 233 properties were generated for them in total. In Devereux et al. (2014), 35 fruit and 29 vegetable SCs were

[26] https://sites.google.com/site/kenmcraelab/norms-data.
[27] https://cslb.psychol.cam.ac.uk/propnorms.
[28] This led to the exclusion of cucumber as a fruit and of parsley as a vegetable, each mentioned by 3 participants in the Devereux et al. dataset. In the McRae et al. data, "a vegetable" was generated by 14 participants for *parsley* and thus included.

identified. Avocado (17F/8V), gherkin (4F/17V), pumpkin (4F/16V), rhubarb (12F/7V) and tomato (22F/5V) are borderline cases. A total of 505 unique properties was available. The amount is more than double of the McRae et al. data, because properties that were only mentioned by two participants were included, compared to 5 in the McRae et al. data. I made the decision to exclude all properties with a PF smaller than 3, which left 362 properties for further analysis. Table 22 shows the number of properties that was produced for each fruit and vegetable SC in both datasets.

Table 22: Number of properties reported in McRae et al. (2005) and Devereux et al. (2012) data for each fruit (left) and vegetable (right) SC.

fruit SC	Properties in McRae et al.	Properties in Devereux et al.
apple	19	21
apricot	NA	19
avocado	10	22
banana	15	24
blueberry	14	24
cantaloupe	14	NA
cherry	12	18
coconut	21	26
cranberry	14	NA
currant	NA	24
dates	NA	23
gherkin	NA	24
grape	18	22
grapefruit	22	29
honeydew	14	NA
kiwi_fruit	NA	30
lemon	17	21
lime	13	22
mandarin	15	NA
melon	NA	25
nectarine	17	22
olive	17	29
orange	16	22
peach	15	19
pear	13	26
pineapple	17	35
plum	15	28
prune	18	25
pumpkin	12	28
raisin	14	23
raspberry	11	21
rhubarb	11	30
satsuma	NA	24
strawberry	17	32
sultana	NA	28
tangerine	16	25
tomato	13	33
watermelon	NA	26

vegetable SC	Properties in McRae et al.	Properties in Devereux et al.
artichoke	NA	22
asparagus	9	24
aubergine	NA	27
avocado	10	22
bean	NA	33
beets	15	NA
broccoli	12	23
brussel_sprouts	NA	29
cabbage	16	21
carrot	18	27
cauliflower	14	22
celery	15	20
corn	13	25
cucumber	16	27
eggplant	9	NA
garlic	NA	27
gherkin	NA	24
leek	NA	28
lettuce	18	19
mushroom	16	28
olive	17	29
onion	16	22
parsley	10	NA
peas	12	19
pickle	18	NA
potato	19	34
pumpkin	12	28
radish	16	28
rhubarb	11	30
spinach	11	20
sweet_potato	NA	25
tomato	13	33
turnip	9	23
yam	7	NA
zucchini	15	22

To reduce the property list to a reasonably testable amount, the following criteria were applied. The amount and percentage of properties deleted due to each criterion are in Table 23. First, all unique properties, those listed for only one SC, were excluded. This was 61% of the properties. Then, the category production frequency (CPF) was determined for each property by counting for how many fruit and vegetable SCs they were generated, excluding borderline cases. All properties with a CPF of less than 3 were excluded[29]. This was 16% of the properties. The resulting list was examined and properties that felt to be too general to have an impact on typicality were excluded[30]. In a last step, the remaining 118 properties were inspected to see if there were overlaps between the datasets. 33 such cases were found, leaving a final list of 85 properties.

Table 23: Number and percentage of excluded properties from the McRae et al. and Devereux et al. property lists.

	McRae et al.		Devereux et al.		total	
before exclusion	233		362		595	
uniqueness	150	64%	210	58%	360	61%
CPF<3	33	14%	62	17%	95	16%
generality	7	3%	15	4%	22	4%
after exclusion	43	18%	75	21%	118	20%
				Overlap	33	
				Total included	85	14%

The formulations of the properties do not always sound natural and the final formulations are based on native English speakers' feedback. All properties together with their CPFs are in Table 84 (appendix).

6.1.3 **Design and procedure**

An online questionnaire was designed using the website Qualtrics. After consenting to the use of their data in anonymised form and for research purposes only and confirming that English is their first language, the question type was introduced to the participants with a short example about land and aquatic animals. The instructions were

> Please imagine that an animal is described to you with one of the properties below. You have to decide how probable it is based on each of these properties that an aquatic and not a land animal is described
>
> 5 means that it is very probably an aquatic animal.
>
> -5 means that it is very probably a land animal.

[29] "grows_on_vines" was wrongfully included despite having CPF=1 for fruit (grape) and CPF=1 for vegetables (cucumber) due to an error in the formula counting its mention for borderline cases (pumpkin, tomato) raising the CPF for vegetable to 3.

[30] "a berry", "is citrus", "grows in Florida", "a fruit", "is edible", "a vegetable", "a root" from McRae et al. and "is a fruit", "does grow", "is eaten/edible", "is a berry", "is food", "is a vegetable", "is grown", "is a plant", "is found in Britain UK England", "is related to cabbages", "is a root vegetable", "is a dried grape", "is a dried fruit", "is a citrus fruit", and "is citrus" from the Devereux et al. data.

> 0 means that both are equally probable based on this property.
>
> Please only select 0 if you see no tendency at all. That is, the property does not discriminate at all between something being an aquatic or land animal.
>
> You can grade your response with the intermediate scale values.

The properties were intuitively chosen to include some that are discriminative for aquatic animals (*has gills*, *swims*) and land animals (*lives in trees*, *speaks English*), one that is not discriminating between the two and should be rated with 0 (*has a liver*, as it applies to most animals) and three that apply to both categories, *lays eggs*, *is yellow* and *is small*, for which people could have differing intuitions.

After the example, the participants were asked to complete an analogous task for *fruit* and *vegetable*. Participants gave their ratings on an 11-point-scale from -5 (very probably a vegetable) to 5 (very probably a fruit) with 0 meaning *fruit and vegetable equally probable*. Only the extremes and 0 were labelled. A screenshot of the scale and response format is in Figure 33. The 85 properties were presented in a randomly ordered list on one page.

> Please decide on the basis of different properties whether a food item is more probably a fruit or a vegetable. Please rate the probability for each of the following properties separately.
>
> 5 means that it is very probable that a food item having this property is a fruit (not a vegetable).
>
> -5 means that a food item having this property is very probably a vegetable.
>
> 0 means that fruit and vegetable are equally probable based on this property.
>
> Please only select 0 if you see no tendency at all. That is, the property does not discriminate at all between a food item being a fruit or a vegetable.
>
> You can grade your response with the intermediate scale values.

Figure 33: Screenshot from the scale and response format of experiment.

The last page of the questionnaire served to collect general demographic data (gender, English dialect, country of residence, level of education), asked participants to self-evaluate if they have knowledge of botany (yes or no) and to indicate their favourite fruit and vegetable.

6.1.4 **Results: Reliability**

The mean completion time was 8:30 minutes (SD = 3:37 minutes). The ratings' intersubjective reliability was assessed with person-group-correlations, i.e., correlations of each participant's ratings with the mean of the remainder of the group (Table 24). The mean is 0.7 (SD = 0.13) with values ranging between 0.36 and 0.81. Cronbach's alpha is .97. To get a more concrete picture of the intersubjectivity of the result compare the means and the SDs of the ratings in Table 85 in the appendix. For example, means around 2 have SDs around 2. This shows that the intersubjectivity is there, but it is not particularly high.

Table 24: Each participants ratings' correlation with the mean of the remainder of the group.

participant	person-group-correlation
1	0.81
2	0.77
3	0.77
4	0.68
5	0.81
6	0.66
7	0.78
8	0.76
9	0.81
10	0.81
11	0.79
12	0.77
13	0.75
14	0.36
15	0.68
16	0.82
17	0.82
18	0.80
19	0.67
20	0.41
21	0.81
22	0.67
23	0.64
24	0.43
25	0.58
26	0.65
27	0.51
28	0.74
29	0.81
30	0.59
Mean	0.70
SD	0.13

6.1.5 **Results: Ratings**

Summary statistics for all properties are in the appendix in Table 85. The goal of this experiment was to find properties that clearly discriminate between fruit and vegetables, which is the case for very high and very low mean ratings. Table 25 and Figure 34 summarise the distribution of ratings and means and medians per rating interval. 50% of all ratings were 0. The mean of 40 properties (47%) and median of 62 (64%) lie between -1 and 1, indicating that participants perceived more than half of properties as non-discriminating. 18 (21%) of means and 10 (12%) of medians are smaller than -1 and thus diagnostic properties for vegetables. Slightly more, 25 (29%) of means and 13 medians (15%), are greater than 1 and diagnostic for fruit.

Table 25: Amount of all ratings, means and medians from experiment 1 in scale intervals.

	-5	-4	-3	-2	-1	0	1	2	3	4	5
ratings	119	90	130	119	103	1287	139	130	163	119	130
%	5%	4%	5%	5%	4%	50%	5%	5%	6%	5%	5%

	[-5,-4)	[-4,-3)	[-3,-2)	[-2,-1]	(-1,1)	[1,2]	(2,3]	(3,4]	(4,5]
mean	0	3	6	9	40	15	10	2	0
%	0%	4%	7%	11%	47%	15%	12%	2%	0%
median	0	3	5	5	54	7	9	1	1
%	0%	4%	6%	2%	64%	2%	11%	1%	1%

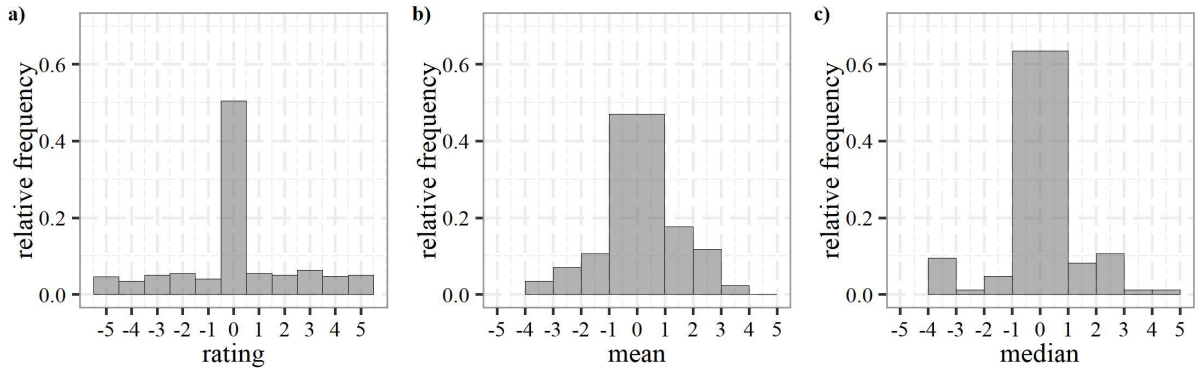

Figure 34: Histograms for a) all ratings (11 bins), b) rating means in intervals (9 bins), c) rating medians in intervals (9 bins) from experiment 1.

Figure 35 and Figure 36 show the histograms for all properties with mean ratings ≥0.3 and <0.3, respectively. Most properties either received ratings between 0 and 5 or between -5 and 0, which shows that the tendency whether the property is more indicative for fruit or vegetable was in general agreed upon between participants and they varied more in the estimation of the degree of diagnosticity. It can also be seen that the mode for most properties is 0.

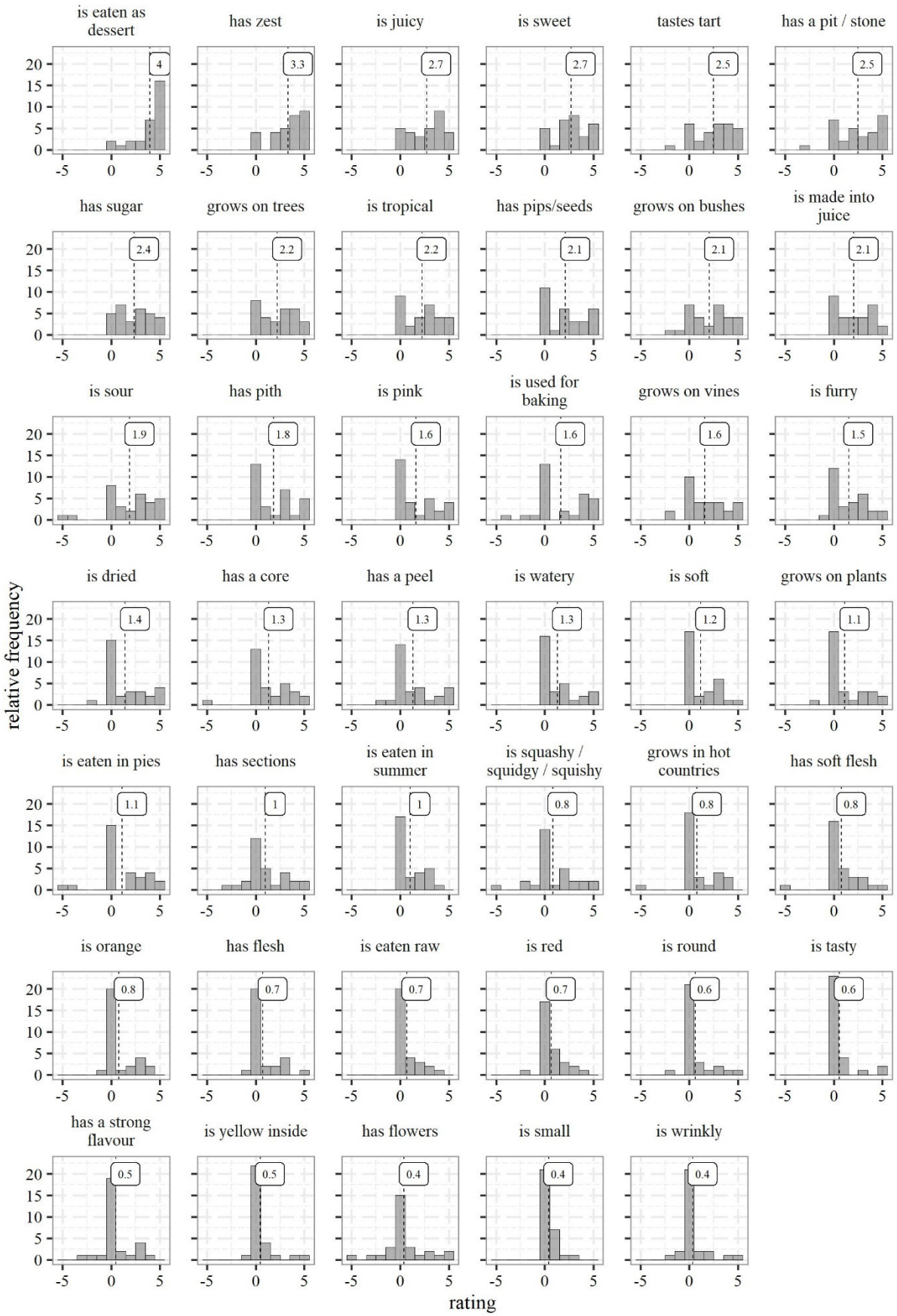

Figure 35: Relative frequencies of ratings per property and their means (vertical line) in experiment 1 for all means ≥0.3.

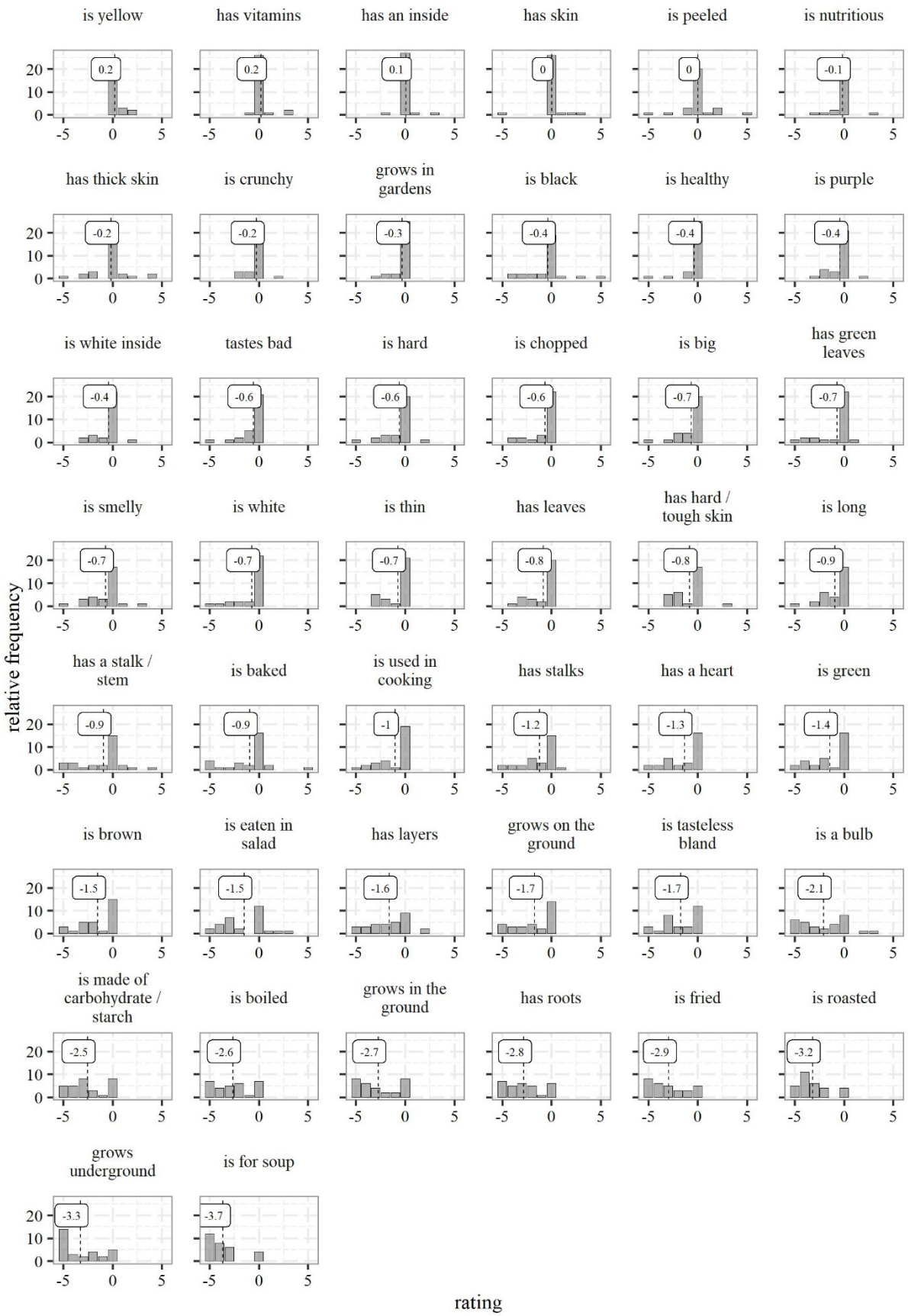

Figure 36: Relative frequencies of ratings per property and their means (vertical line) in experiment 1 for all means <0.3.

6.1.6 **Discussion**

This experiment shows that half of the properties that were identified to be associated with *fruit* and *vegetable* are not perceived to be diagnostic for either category. 50% of properties had ratings with no clear tendency, 21% have a diagnostic tendency for *vegetable* and 29% for *fruit*. The goal to narrow down the large list of properties to few which are diagnostic was achieved: almost half of the properties could be eliminated. The elimination of non-diagnostic properties (diagnosticity values close to 0) is justified insofar they do not contribute significantly to the typicality of a SC or a C, and their elimination reduces procedural complexity.

6.2 **Experiment 2**

The aim of the second experiment was to gather data on the subjective probability of the properties showing a diagnostic tendency from experiment 1, the reversed probability Pr (P|C). The question was how participants estimate the probability with which the fruit and vegetable category have the properties in question. The most probable and at the same time most diagnostic properties were then used in experiment 3.

6.2.1 **Participants**

Participants were 30 paid volunteers (16 female, 14 male, mean age 32.1, SD = 12.1) recruited via the platform Prolific. The demographic information is summarised in Table 83 in the appendix. 19 were from the UK, 7 from the USA, 2 from Ireland and one each from Canada and Australia. 21 of them characterised their English as British English and 9 as American. All participants were at least high school graduates, most had done some college (8) or a bachelor (8). 4 of them stated that they had some knowledge of botany and 26 that they had not.

6.2.2 **Stimuli**

Stimuli were those 45 properties from experiment 1 that showed a slight diagnostic tendency for either fruit or vegetables, which was taken to be the case when their mean rating was greater than or equal to 1 or smaller than or equal to -1. The properties and their mean ratings from experiment 1 are in Table 26.

Table 26: Properties from experiment 1 with a mean diagnosticity greater or equal 1 or smaller or equal -1.

property	mean	property	mean
is eaten as dessert	4.0	grows on plants	1.1
has zest	3.3	is eaten in pies	1.1
is sweet	2.7	is eaten in summer	1.0
is juicy	2.7	has sections	1.0
tastes tart	2.5	is used in cooking	-1.0
has a pit/stone	2.5	has stalks	-1.2
has sugar	2.4	has a heart	-1.3
is tropical	2.2	is green	-1.4
grows on trees	2.2	is brown	-1.5
has pips/seeds	2.1	is eaten in salad	-1.5
grows on bushes	2.1	has layers	-1.6
is made into juice	2.1	grows on the ground	-1.7
is sour	1.9	is tasteless/bland	-1.7
has pith	1.8	is a bulb	-2.1
is pink	1.6	is made of carbohydrate/starch	-2.5
is used for baking	1.6	is boiled	-2.6
grows on vines	1.6	grows in the ground	-2.7
is furry	1.5	has roots	-2.8
is dried	1.4	is fried	-2.9
has a core	1.3	is roasted	-3.2
is watery	1.3	grows underground	-3.3
has a peel	1.3	is for soup	-3.7
is soft	1.2		

6.2.3 **Design and procedure**

The questionnaire was designed on the website www.soscisurvey.de. After giving informed consent to the use of the data and the confirmation that all participants' first language is English, a short example about either aquatic or land animals was shown to familiarise the participants with the question type. Two clearly impossible properties (*has gills* for land animals and *speaks English* for aquatic animals) were included.

> Our survey is about explaining the meaning of words with probabilities of properties. We will begin with a short example.
>
> Below you find a list of properties. Please imagine that you have to explain to someone based on these properties what a land animal is. In order to do this, please rate the probability of a land animal having the following properties on a scale from 0 to 5.
>
> 0 means that it is extremely improbable that a land animal has this property.
>
> 5 means that it is extremely probable that a land animal has this property.
>
> You can use the numbers in between to grade your responses.
>
> Please answer intuitively.
>
> If you know something is a land animal, then how likely is it to have each property?

Afterwards, the probability of the 45 selected properties was rated for fruit and vegetables on separate pages in a random order. Answers were given on a 6-point-scale with labelled

extremes, where 0 was labelled “extremely improbable” and 5 “extremely probable”. A screenshot of the scale and response format is in Figure 37. The concrete instructions were:

> Please imagine that you have to explain to someone what a fruit is. To do this, please rate on a scale from 0 to 5 how probable it is that fruit have the following properties.
>
> 0 means that it is extremely improbable that fruit have this property.
>
> 5 means that it is extremely probable that fruit have this property.
>
> You can use the numbers in between to grade your responses.
>
> If you know something is a fruit, then how likely is it to have each property?

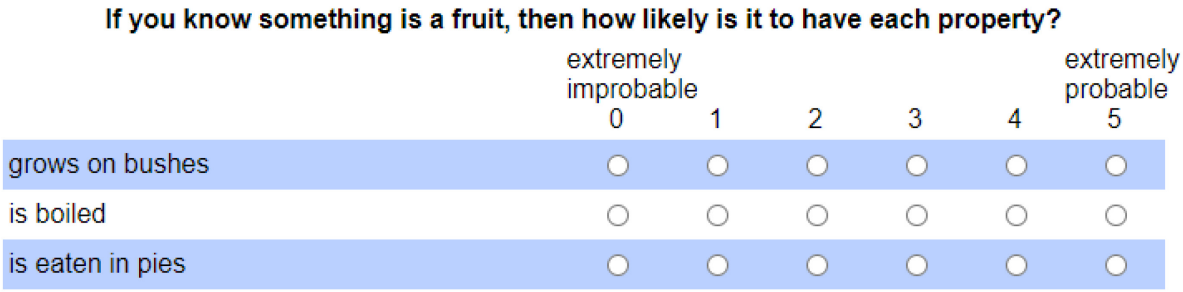

Figure 37: Screenshot of the answer format of experiment 2. The answer format for vegetable was identical.

The last page of the questionnaire collected general demographic data (gender, English dialect, country of residence, level of education), asked participants to self-evaluate if they have knowledge of botany (yes or no) and to indicate their favourite fruit and vegetable.

6.2.4 **Results: Reliability**

The mean completion time was 8:06 minutes (SD = 3:03 minutes) after exclusion of one outlier who took 2 hours and 24 minutes.

The correlations of each participant with the remainder of the group are in Table 27. The mean is 0.69 (SD = 0.09) with a minimum of 0.41 and maximum of 0.86. There is no considerable difference in the correlation for only fruit (mean = 0.71, SD = 0.11) or vegetable (mean = 0.66, SD = 0.13) ratings.

The means and SDs in Table 86 in the appendix show that the SDs are generally <1.5 and thus intersubjective agreement for most properties.

Table 27: Correlation of each subject's ratings with the mean of the remainder of the group for experiment 2 for all ratings and split by fruit and vegetable ratings.

subject	correlation with remainder	only fruit ratings	only vegetable ratings	diff
1	0.86	0.89	0.85	0.04
2	0.65	0.84	0.38	0.47
3	0.82	0.81	0.83	0.02
4	0.62	0.69	0.52	0.17
5	0.68	0.69	0.67	0.02
6	0.70	0.72	0.68	0.04
7	0.69	0.61	0.76	0.16
8	0.78	0.82	0.73	0.08
9	0.80	0.79	0.79	0.00
10	0.74	0.75	0.73	0.02
11	0.70	0.76	0.64	0.12
12	0.70	0.65	0.74	0.09
13	0.74	0.67	0.81	0.14
14	0.55	0.74	0.31	0.43
15	0.57	0.52	0.61	0.09
16	0.71	0.75	0.68	0.06
17	0.67	0.65	0.66	0.02
18	0.41	0.35	0.47	0.12
19	0.68	0.70	0.67	0.03
20	0.70	0.71	0.78	0.07
21	0.76	0.81	0.70	0.11
22	0.79	0.79	0.76	0.03
23	0.68	0.79	0.63	0.16
24	0.64	0.55	0.72	0.17
25	0.67	0.72	0.61	0.10
26	0.66	0.79	0.53	0.25
27	0.67	0.72	0.65	0.06
28	0.81	0.81	0.80	0.01
29	0.61	0.63	0.63	0.00
30	0.60	0.64	0.54	0.10
Mean	0.69	0.71	0.66	0.11
SD	0.09	0.11	0.13	0.11

6.2.5 **Results: Ratings**

The means and SDs for all properties are in Table 86 in the appendix. The frequency distributions of ratings and means are in Table 28. For both categories, 0 is the least frequent rating with 10% and 14%. The most frequent rating for *fruit* is 5 with 26% and for *vegetable* 1, 3 and 5 all received 18% of ratings. Most means for *fruit* are between 3 and 4 (16 properties, 36%) and for *vegetable* between 2 and 3 (13, 29%). Between 1 and 2 and between 3 and 4 there are 27% resp. 24%. For both categories, only 3 properties (7%) have a mean rating smaller than 1.

Table 28: Frequency distributions of means for fruit and vegetables in experiment 2.

		0	1	2	3	4	5
ratings	fruit	138	201	178	241	246	346
		10%	15%	13%	18%	18%	26%
	vegetables	192	247	214	240	213	244
		14%	18%	16%	18%	16%	18%

		[0,1)	[1,2)	[2,3)	[3,4)	[4,5]
means	fruit	3	8	10	16	8
		7%	18%	22%	36%	18%
	vegetables	3	12	13	11	6
		7%	27%	29%	24%	13%

The ratings are visualised ordered by the diagnosticity ratings from experiment 1 in Table 29. From the 8 properties with a mean ≥4 for fruit, 7 have a high diagnosticity (mean > 2) as well. The property *is eaten in summer* has a high mean probability rating for fruit (4.4), but a low mean diagnosticity rating of 1. From the 6 properties with a high mean for vegetables, 5 have a high diagnosticity (mean < -2). The property *is used in cooking* has a high mean probability rating for vegetable (4.8), but a low diagnosticity rating of -1.

Table 29: Mean ratings from experiment 2, their difference and mean ratings from experiment 1.

property	mean Pr(P|fruit) rating	mean Pr(P|vegetable) rating	difference	mean Pr(fruit|P) rating
is eaten as dessert	4.2	0.8	3.4	4.0
has zest	3.5	1.0	2.4	3.3
is sweet	4.4	1.8	2.6	2.7
is juicy	4.4	1.7	2.7	2.7
tastes tart	3.0	1.4	1.6	2.5
has a pit/stone	3.7	1.1	2.6	2.5
has sugar	4.6	2.1	2.5	2.4
is tropical	3.9	1.7	2.3	2.2
grows on trees	4.3	1.6	2.7	2.2
has pips/seeds	4.4	2.2	2.2	2.1
is made into juice	4.7	2.7	1.9	2.1
grows on bushes	3.9	1.9	2.0	2.1
is sour	2.7	1.5	1.3	1.9
has pith	3.1	1.3	1.8	1.8
is used for baking	3.6	2.2	1.5	1.6
is pink	2.9	0.9	1.9	1.6
grows on vines	3.2	2.1	1.2	1.6
is furry	2.7	0.9	1.8	1.5
is dried	3.6	1.9	1.8	1.4
has a core	3.8	2.1	1.7	1.3
has a peel	3.9	2.9	1.0	1.3
is watery	3.6	2.3	1.3	1.3
is soft	3.7	2.3	1.4	1.2
grows on plants	3.6	3.4	0.2	1.1
is eaten in pies	3.6	2.7	0.9	1.1
is eaten in summer	4.4	3.3	1.1	1.0
has sections	3.5	2.2	1.3	1.0
is used in cooking	3.6	4.8	1.2	-1.0
has stalks	2.2	3.3	1.2	-1.2
has a heart	0.9	1.5	0.7	-1.3
is green	2.6	3.9	1.3	-1.4
is brown	1.6	3.0	1.4	-1.5
is eaten in salad	2.5	3.9	1.4	-1.5
has layers	2.4	3.2	0.8	-1.6
grows on the ground	2.3	3.8	1.5	-1.7
is tasteless/bland	0.9	2.1	1.2	-1.7
is a bulb	1.1	2.5	1.4	-2.1
is made of carbohydrate/starch	2.5	3.6	1.1	-2.5
is boiled	1.5	4.1	2.5	-2.6
grows in the ground	1.8	4.1	2.3	-2.7
has roots	1.7	4.1	2.3	-2.8
is fried	1.6	3.3	1.7	-2.9
is roasted	1.3	4.1	2.9	-3.2
grows underground	0.6	3.6	3.0	-3.3
is for soup	1.0	4.4	3.4	-3.7

There is a remarkable correspondence between the mean diagnosticity ratings from experiment 1 and the difference between the mean probabilities for *fruit* and *vegetable* from experiment 2, visualised in Figure 38. The correlation is .98. This is also interesting in view of the fact that the diagnostic ratings were a contrastive comparison – the contrast category was explicitly included in the scale – while the probability ratings were not contrastive and did not mention the contrast category in the scale.

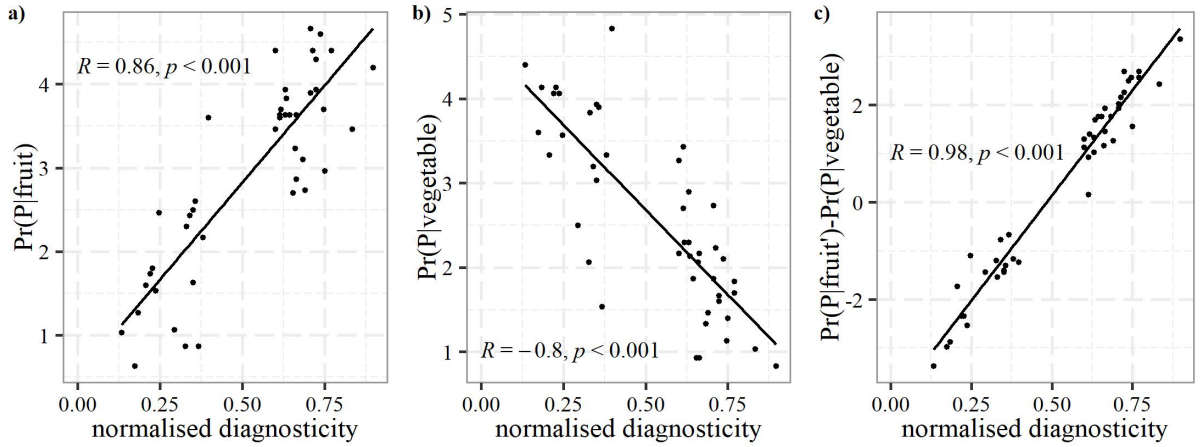

Figure 38: Normalised diagnosticity from experiment 1 against a) mean rated probability for fruit, b) mean rated probability for vegetable and c) the difference between fruit and vegetable probability ratings.

6.2.6 **Discussion**

It was possible to identify properties for both categories that are diagnostic as well as frequent. Diagnosticity and the difference between *fruit* and *vegetable* probabilities have an almost perfect correlation – as they should if human probability judgments follow the laws of probability. An application of Bayes' theorem to the data from experiment 1 and 2 is in section 6.4.3.

6.3 **Experiment 3**

Experiment 3 completes the data required to test the probabilistic prototype formula for typicality prediction by collecting two data types for *fruit* and *vegetable* SCs: typicality ratings of SCs for Cs, and probability ratings of the previously identified diagnostic and frequent properties of C for SCs, Pr (P|SC). The data from this experiment makes a comparison between C and SC probabilities possible.

6.3.1 **Participants**

Participants were 62 paid volunteers hired via Prolific. They completed the online questionnaire for either *fruit* (30) or *vegetable* (32). Seven participants completed both the *fruit* and the *vegetable* questionnaire. The demographic information is summarised in Table 83 in the appendix. The mean age was 33 for fruit (SD = 13) and 32.6 for vegetables (SD = 12.6). Two thirds of the participants were between 18 and 30. Both studies had a gender ratio of 2:1 for female vs. male and participants were predominantly residing in the UK and characterised their English as British dialect. The educational level was mixed but all except one participant completed high school. Around 80% of participants stated that they had no knowledge of botany.

Table 30: Age, gender, English dialect, country of residence, level of education and botany knowledge of participants in experiment 3.

		fruit	vegetables
Gender	male	10	9
	female	20	21
	other	0	2
Age	[18,30)	15	17
	[30,40)	6	5
	[40,50)	4	7
	[50,60)	3	1
	[60,70)	2	2
Country of residence	UK	23	18
	USA	5	7
	Canada	2	4
	Australia	0	2
	Spain	0	1
English dialect	British	25	19
	American	4	9
	Other	1	4
Education	Less than high school	0	1
	High school graduate	6	8
	Some college	6	9
	2 year degree	2	1
	4 year degree	0	3
	Professional degree	4	0
	Bachelor	7	6
	Master	2	4
	Doctorate	3	0
Botany knowledge	Yes	5	7
	No	25	25

6.3.2 **Stimuli**

40 SCs were used in total. They are in Table 31 for *fruit* and Table 32 for *vegetable* along with the selection criteria. 18 SCs were uniquely selected for each category and 4 were used in both experiments as borderline cases: *avocado*, *pumpkin*, *tomato* and *rhubarb*. The criterion for the selection was that six each were found in the high, medium and low range of the normalised means from the typicality meta-analysis described in section 3.2. In addition, all chosen SCs fulfilled at least one of the following additional criteria: SCs had a low SD between studies (<0.1), were used in many studies (>7 for *vegetable* and >8 for *fruit*) or were named favourite in the two former experiments (>2 participants mentioned them). The inclusion of borderline

cases leads to four SCs more on the low typicality level for *fruit* and for *vegetable* three more on the low and one (*tomato*) on the medium level.

Table 31: Selected subcategories for fruit (left) with mean normalised typicality from meta-analysis, SD, number of means available from typicality meta-analysis and selection criteria.

subcategory	available means	normalised meta mean	normalised meta SD	SD<0.1	used in >8 studies	borderline	favourite	Sum of criteria
apple	16	0.97	0.03	1	1	0	11	3
banana	14	0.92	0.05	1	1	0	7	3
strawberry	17	0.88	0.07	1	1	0	7	3
peach	15	0.86	0.08	1	1	0	1	2
pineapple	13	0.84	0.08	1	1	0	2	2
grape	15	0.83	0.09	1	1	0	3	3
plum	12	0.81	0.08	1	1	0	1	2
blackberry	8	0.75	0.06	1	0	0	2	1
watermelon	11	0.71	0.12	0	1	0	3	2
passion fruit	5	0.68	0.06	1	0	0	0	1
blueberry	11	0.68	0.12	0	1	0	2	1
mango	12	0.68	0.12	0	1	0	7	2
lime	11	0.65	0.11	0	1	0	0	1
papaya	9	0.60	0.14	0	1	0	0	1
pomegranate	12	0.57	0.16	0	1	0	0	1
fig	13	0.52	0.12	0	1	0	0	1
prune	6	0.53	0.07	1	0	0	0	1
coconut	15	0.51	0.15	0	1	0	0	1
rhubarb	2	0.45	0.04	1	0	1	0	2
avocado	11	0.43	0.13	0	1	1	1	2
tomato	10	0.37	0.12	0	1	1	1	2
pumpkin	10	0.32	0.14	0	1	1	0	2

Table 32: Selected subcategories for vegetables with mean normalised typicality from meta-analysis, SD, number of means available from the typicality meta-analysis and selection criteria.

subcategory	available means	normalised meta mean	normalised meta SD	SD<0.1	used in >7 studies	borderline	favorite	sum of criteria
carrot	12	0.90	0.07	1	1	0	8	3
pea	10	0.85	0.08	1	1	0	1	2
lettuce	13	0.85	0.07	1	1	0	0	2
spinach	12	0.84	0.09	1	1	0	2	2
cauliflower	11	0.82	0.11	0	1	0	0	1
broccoli	7	0.79	0.13	0	0	0	8	1
zucchini	7	0.78	0.04	1	0	0	3	2
tomato	12	0.73	0.13	0	1	1	0	2
potato	13	0.71	0.17	0	1	0	7	2
radish	10	0.70	0.10	1	1	0	0	2
green onion	3	0.70	0.03	1	0	0	0	1
corn	8	0.69	0.17	0	1	0	0	1
onion	12	0.69	0.17	0	1	0	5	2
eggplant	10	0.69	0.12	0	1	0	0	1
mushroom	8	0.62	0.08	1	1	0	1	2
pumpkin	5	0.58	0.13	0	0	1	0	1
parsley	10	0.57	0.10	1	1	0	0	2
rhubarb	3	0.53	0.11	0	0	1	0	1
sweet potato	2	0.52	0.11	0	0	0	6	1
avocado	3	0.51	0.11	0	0	1	0	1
pickle	8	0.50	0.08	1	1	0	0	2
garlic	9	0.41	0.11	0	1	0	1	1

The properties were selected based on the results of the former two experiments. The mean ratings from experiment 1 were normalised with the formula $\frac{\Pr(C|P)+5}{10}$ to lie between 0 and 1. A high normalised mean corresponds to a high diagnosticity for fruit and a low normalised mean corresponds to a high diagnosticity for vegetable. The ratings from experiment 2 were normalised to lie between 0 and 1 by dividing them by 5. All properties that had a normalised mean $\Pr(C|P)$ rating greater than 0.7 for fruit and smaller than 0.3 for vegetable, and at the same time a normalised probability rating greater than 0.8 for either fruit or vegetable were used in the final experiment. 4 control properties with either high diagnosticity but low probability (*tastes tart* for fruit and *has a heart* for vegetables) or high probability and low diagnosticity (*grew on a plant* for fruit and *will be used in cooking* for vegetables) were included. After a pretest of the questionnaire revealed uncertainty of the participants in interpreting the question, the phrasing of the properties was adjusted to past tense for growing conditions and future tense for preparation methods and common dishes to make clear that participants are asked to rate the probability of this property for one random member of the SC in question. Table 33 shows the selected properties, their new phrasing and the normalised mean ratings from experiment 1 and 2.

Table 33: Selected properties for experiment 3 with new phrasing and normalised mean ratings from experiment 1 and 2.

property	new phrasing	Mean Pr(fruit|P) norm	Mean Pr(P|fruit)	Mean Pr(P|vegetables)
is eaten as dessert	will be eaten as dessert	0.90	0.84	0.17
is sweet		0.77	0.88	0.37
is juicy		0.77	0.88	0.34
tastes tart		0.75	0.59	0.28
has sugar		0.74	0.92	0.42
grows on trees	grew on trees	0.72	0.86	0.32
has pips/seeds	has pips/seeds	0.71	0.88	0.45
is made into juice	will be made into juice	0.71	0.93	0.55
grows on plants	grew on a plant	0.61	0.72	0.69
is used in cooking	will be used in cooking	0.40	0.72	0.97
has a heart		0.37	0.17	0.31
is boiled	will be boiled	0.24	0.31	0.81
grows in the ground	grew in the ground	0.23	0.36	0.83
has roots		0.22	0.35	0.81
is roasted	will be roasted	0.18	0.25	0.83
is for soup	will be eaten in soup	0.13	0.21	0.88

6.3.3 **Design and procedure**

Experiment 3 comprised two structurally identical questionnaires: one for *fruit* and one for *vegetable*. They had two parts: typicality ratings of SCs for Cs, and property probability ratings for SCs. After consenting to the use of their data in anonymised form and solely for research purposes and confirming that English was their first language, participants were shown a modified version of the instructions used in Rosch (1975b) to introduce them to typicality.

Table 34 shows the original version and the modified version with all modifications underlined. It differs in three ways. First, instead of referring to SCs as *category members*, they are referred to as *food items* to avoid the implication that all SCs are definite members of the categories. Second, Rosch's example for *apple* in the category *fruit* was changed to *armchair* in the category *furniture*, to avoid making statements about *fruit* before participants had completed the task. Third, the option to mark SCs as non-members was included, but participants were asked to rate typicality for non-members as well. As this was found to be counter-intuitive by some participants in the pretest, the example that *dolphin* is a better example for *fish* than *chair*, even if they are not, technically, fish, was added.

Table 34: Instruction used in (Rosch 1975b, p. 198) and modified version with underlined modifications.

Rosch (1975)	Modified version
This study has to do with what we have in mind when we use words which refer to categories. Let's take the word red as an example. Close your eyes and imagine a true red. Now imagine an orangish red ... imagine a purple red. Although you might still name the orange red or the purple red with the term red, they are not as good examples of red (as clear cases of what red refers to) as the clear "true" red. In short, some reds are redder than others. The same is true for other kinds of categories. Think of dogs. You all have some notion of what a 'real dog', a 'doggy dog' is. To me a retriever or a German shepherd is a very doggy dog while a Pekinese is a less doggy dog. Notice that this kind of judgment has nothing to do with how well you like the thing; you can like a purple red better than a true red but still recognize that the color you like is not a true red. You may prefer to own a Pekinese without thinking that it is the breed that best represents what people mean by dogginess.	This study has to do with what we have in mind when we use words which refer to categories. Let's take the word red as an example. Close your eyes and imagine a true red. Now imagine an orangish red ... imagine a purple red. Although you might still name the orange red or the purple red with the term red, they are not as good examples of red (as clear cases of what red refers to) as the clear "true" red. In short, some reds are redder than others. The same is true for other kinds of categories. Think of dogs. You all have some notion of what a 'real dog,' a 'doggy dog' is. To me a retriever or a German shepherd is a very doggy dog while a Pekinese is a less doggy dog. Notice that this kind of judgment has nothing to do with how well you like the thing; you can like a purple red better than a true red but still recognize that the color you like is not a true red. You may prefer to own a Pekinese without thinking that it is the breed that best represents what people mean by dogginess.
On this form you are asked to judge how good an example of a category various instances of the category are. At the top of the page is the name of a category. Under it are the names of some members of the category. After each member is a blank. You are to rate how good an example of the category each member is on a 7-point scale. A 1 means that you feel the member is a very good example of your idea of what the category is. A 7 means you feel the member fits very poorly with your idea or image of the category (or is not a member at all). A 4 means you feel the member fits moderately well. For example, one of the members of the category *fruit* is *apple*. If *apple* fit well your idea or image of *fruit*, you would put a 1 after it; if *apple* fit your idea of *fruit* very poorly you would put a 7 after it; a 4 would indicate moderate fit. Use the other numbers of the 7-point scale to indicate intermediate judgments. Don't worry about why you feel that something is or isn't a good example of the category. And don't worry about whether it's just you or people in general who feel that way. Just mark it the way you see it.	On the next two pages you are asked to judge how good an example of fruit (vegetables) various food items are. At the top of the page is the name of a category. Under it are the names of some food items. After each member is a scale. You are to rate how good an example of the category each food item is on a 7-point scale. A 1 means that you feel the item is a very good example of your idea of fruit (vegetables). A 7 means you feel the item fits very poorly with your idea. A 4 means you feel the item fits moderately well. For example, an item related to the category *furniture* is *armchair*. If *armchair* fit well your idea or image of *furniture*, you would put a 1 after it; if *armchair* fit your idea of *furniture* very poorly you would put a 7 after it; a 4 would indicate moderate fit. Use the other numbers of the 7-point scale to indicate intermediate judgments. If you think that the food item is not a fruit or vegetable at all, please indicate this by ticking the corresponding box, but nevertheless rate it. It is possible for objects to be better or worse examples for categories even if they are not, technically, members. For example: dolphins are better examples for the category fish than chairs. Don't worry about why you feel that something is or isn't a good example of the category. And don't worry about whether it's just you or people in general who feel that way. Just mark it the way you see it.

The next page showed a list of 22 SCs in randomised order and a 7-point rating scale on the right. Figure 39 is a screenshot of the response format. The extremes were labelled "very good example" and "very bad example" respectively. Next to the scale was a box labelled "not a fruit" or "not a vegetable" which could be checked optionally.

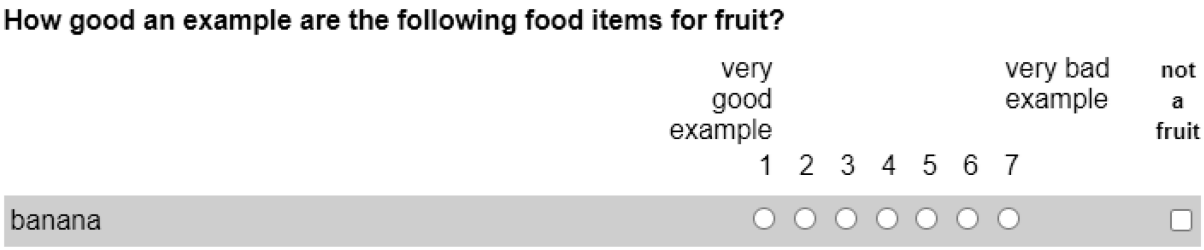

Figure 39: Screenshot of the response format for typicality ratings in experiment 3.

Then, the second part about property probability ratings was introduced with an example for land animals. The properties were the same as in the second experiment, but the response format was changed to a slider scale with 101 points from 0 to 100 (normalised to lie between 0 and 1 by dividing by 100) with the extremes labelled "extremely improbable" and "extremely probable" and a wedge as graphical anchor. A screenshot of the response format is in Figure 40. The instructions were:

> The second part of our survey is about explaining the meaning of words with probabilities of properties. We will begin with a short example.
>
> Below you find a list of properties. Please imagine that you have to explain to someone what a land animal is, based on these properties. In order to do so, please rate the probability of a land animal having the following properties on a slider.
>
> Move the slider from the box on the scale. Place it the more to the right, the more probable or frequent you think it is that a random land animal has this property. Place it the more to the left, the less probable or frequent you think it is that a land animal has this property.
>
> Place the slider on the right-most point if you think that any land animal certainly has this property (in other words, certainty means maximal probability). Place the slider on the left-most point if you think that certainly no land animal has this property.
>
> The medium position means that you think that approximately half (50%) of land animals have the property in question.
>
> You can use the whole scale to make fine-grained, intermediate judgements, wherever this makes sense. We ask for your subjective estimation of these frequencies, so do not worry, if you are not sure about some ratings. Just let us know your intuition.

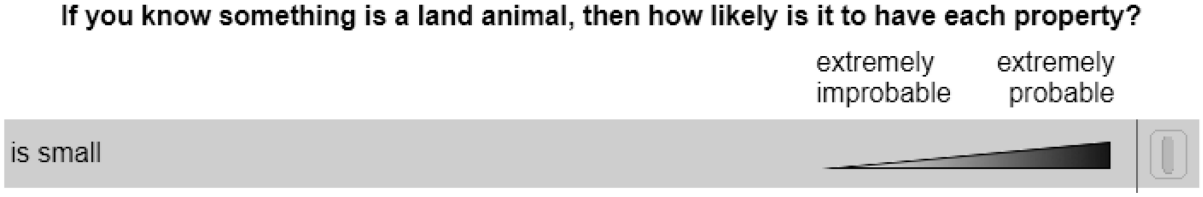

Figure 40: Example of slider response format used in experiment 3.

Then, the participants rated the probability of the selected properties for the 22 SCs, both in randomised order. The concrete instructions, exemplarily for carrots, were:

> Please imagine that you have to explain to someone what a carrot is. To do this, please rate on a slider how probable or frequent it is that carrots have the following properties.
> Move the slider from the box on the scale. Place it the more to the right, the more probable or frequent you think it is that a random carrot has this property. Place it the more to the left, the less probable or frequent you think it is that a random carrot has this property.
> Place the slider on the right-most point if you think that any carrot certainly has this property (in other words, certainty means maximal probability). Place the slider on the left-most point if you think that certainly no carrot has this property.
> The medium position means that you think that approximately half (50%) of carrots have the property in question.

> You can use the whole scale to make fine-grained, intermediate judgements, wherever this makes sense. We ask for your subjective estimation of these frequencies, so if you are not sure about some ratings, do not worry. Just let us know your intuition.

The last page of the questionnaire served to collect general demographic data (gender, English dialect, country of residence, level of education), asked participants to self-evaluate if they have knowledge of botany (yes or no) and to indicate their favourite fruit and vegetable. The questionnaire contained for each participant 22 typicality ratings and 352 property probability ratings (16 properties for 22 SCs). It was the longest questionnaire in the series of experiments.

6.3.4 Results: Typicality ratings

Typicality was rated by 30 (fruit) and 32 (vegetables) participants on a 7-point-scale (1 = very good example, 7 = very poor example) for 22 fruit or 22 vegetable SCs.

The mean completion time of the rating task for fruit SCs was 1:30 minutes (SD = 0:41 minutes) after exclusion of two outliers who took more than 5 minutes. Participants took in general time to read the rather long instructions (mean = 59 seconds, SD = 40 seconds), after exclusion of one participant who had completed the vegetable questionnaire before and thus knew the instructions. 3 participants took less than 10 seconds.

The mean completion time for vegetable SCs was 1:47 minutes (SD = 0:47 minutes) after exclusion of two outliers who took more than 5 minutes. Participants took time to read the instructions with a mean completion time of 1:13 minutes (SD = 0:59 minutes), after exclusion of two outliers who took more than 5 minutes and 5 participants who completed the fruit questionnaire before. 3 participants took less than 10 seconds.

6.3.4.1 Reliability

There are two ways to determine the reliability of the typicality ratings: internal, by looking at the person-group-correlations between participants, and external, by comparing the observed mean ratings with those from the literature. Both will be discussed in turn. The intersubjective reliability in terms of means and SDs is discussed in section 6.3.4.2.

The person-group-correlations are in Table 36. The mean is .65 for fruit and .56 for vegetable. With a sample size of 30, correlations of .45 are significant at the .01 level. There are 5 participants for fruit and 8 for vegetable that have correlations < .45 and they were excluded for the further use of the data. The results of this removal are presented in the results section (section 6.3.4.2). After exclusion, the mean person-group-correlations are .77 for fruit and .64 for vegetable.

Table 35: Person-group-correlations of typicality ratings from experiment 3 for a) fruit and b) vegetable.

a) fruit

participant	correlation with remainder
1	.84
2	.86
3	.67
4	.79
5	.93
6	.51
7	.54
8	.42
9	.73
10	.89
11	-.64
12	.35
13	.84
14	.95
15	.73
16	.72
17	.85
18	.16
19	.71
20	.91
21	.81
22	.88
23	.83
24	.92
25	.8
26	.17
27	.66
28	.84
29	.74
30	.68
mean	.67
SD	.32

b) vegetable

participant	correlation with remainder
1	.78
2	.59
3	.81
4	.48
5	.74
6	.75
7	.85
8	.56
9	.74
10	.37
11	.14
12	.43
13	.53
14	.45
15	.6
16	.82
17	.77
18	.71
19	.45
20	.34
21	.48
22	.78
23	.79
24	.18
25	.29
26	.83
27	.68
28	.36
29	.73
30	-.36
31	.59
32	.7
mean	.56
SD	.26

With regard to external reliability, Figure 41 shows the normalised mean ratings, the means from the meta-analysis, the meta means and the difference between mean ratings and meta means. The correlations between the mean ratings from experiment 3 with the meta means are .9 for *fruit* and .8 for *vegetable*. The lower correlation for *vegetable* is mainly due to different means of four SCs: *sweet potato* (.75 vs. .52), *tomato* (.43 vs. .73), *rhubarb* (.28 vs. .53) and *parsley* (.26 vs .57). Without them, the correlation is .86. The by far highest difference for *fruit*

is for *pomegranate* (.89 vs. .57). *mango, passion fruit, blueberry, blackberry, lime, papaya* and *prune* all have means higher than the meta means.

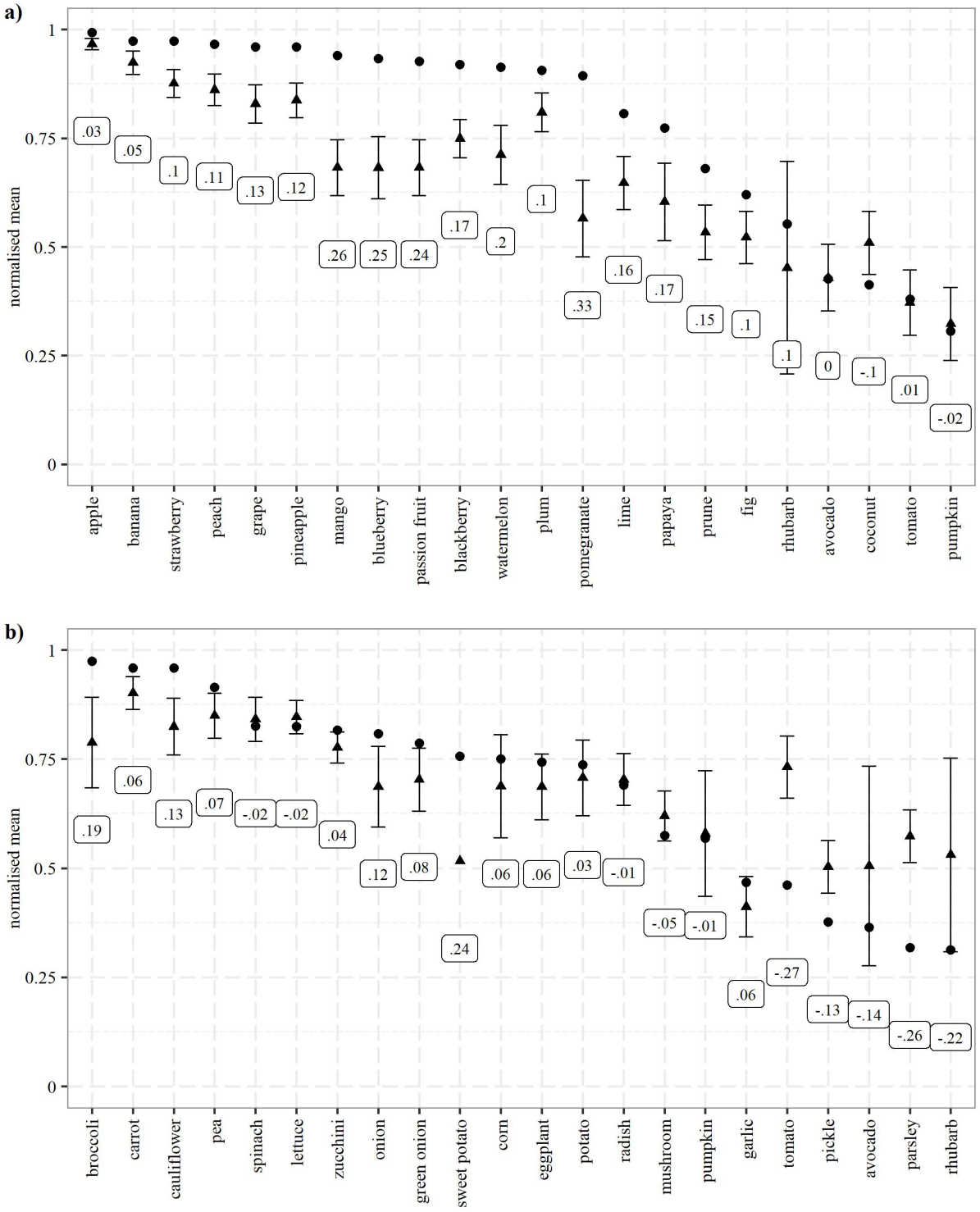

Figure 41: Means from typicality meta-analysis (black triangles with 95% confidence intervals) and means from experiment 3 (black dots) and their differences (boxes below) for a) fruit and b) vegetable SCs.

In addition, our results are similar to those from the studies on which the meta-analysis was based: correlations are in Table 36. The mean correlations with previous studies are .82 (SD = .09) for *fruit* and .65 (SD = .15) for *vegetable*. The lowest correlations for both categories

are with the data from Moreno-Martínez et al. (2014), which was also found to have low correlations with other datasets in the meta-analysis.

Table 36:Pearson correlations of experiment 3 mean typicality ratings with other studies.

	fruit		vegetable	
study	n	r	n	r
Rosch 1975	21	.91	21	.77
McCloskey et al. 1978	13	.9	13	.73
Uyeda et al. 1980	17	.84	13	.64
Hampton et al. 1983	18	.8	17	.88
Malt et al. 1982	9	.93		
Malt et al. 1984	11	.79		
Barsalou 1985	10	.78	12	.58
Brown et al. 1986	18	.79	18	.59
Schwanenflugel et al. 1986 ENG	15	.85		
Schwanenflugel et al. 1986 SPA	15	.72		
Smith et al. 1988	11	.97	10	.83
Ruts et al. 2004	16	.87	16	.73
De Deyne et al. 2008 TYP	16	.76	16	.61
De Deyne et al. 2008 GDN	16	.86	16	.58
Schröder et al. 2012	16	.73	14	.55
Moreno-Martinez et al. 2014	17	.57	17	.3
Hampton FRE	9	.81	NA	NA
mean		.82		.65
SD		.09		.15

6.3.4.2 Ratings

The means and SDs of the *fruit* and *vegetable* typicality ratings before and after cleaning as well as the number of participants who checked the box "not a category member" are in Table 87 in the appendix. The SC selection was supposed to cover the whole typicality spectrum. 6 SCs each have a high, medium and low typicality in the meta-analysis (section 3.2) and 4 SCs were selected as borderline cases. Compared to this expectation, for *fruit* there are more SCs with a mean on the high typicality level: 13 SCs had a mean smaller than 2. Only 3 SCs were rated to be medium typical with a mean rating between 2 and 3 and 6 to be rather untypical with a mean rating greater than 3. For *vegetable*, only 4 SCs have a mean rating smaller than 2. Most SCs are on the medium typicality level: 10 SCs have a mean between 2 and 3. The lower typicality level is covered by 8 SCs with a mean rating greater than 3.

Figure 42 shows the mean and 95% confidence interval after cleaning, i.e., removing all participants with a person-group-correlation < .45. For *fruit*, the SDs after cleaning are much smaller (approximately -1) for SCs with a typicality <2, but almost unchanged for the others. For *vegetable*, the same trend can be seen. For the 4 SCs with a typicality <2, the SD has been almost reduced by half as a result of the cleaning. In general, cleaning has a smaller influence here. Most SCs have a mean typicality >2. The mean SD went from 1.7 to 1.2 for *fruit* and from 1.9 to 1.7 for *vegetable* after cleaning.

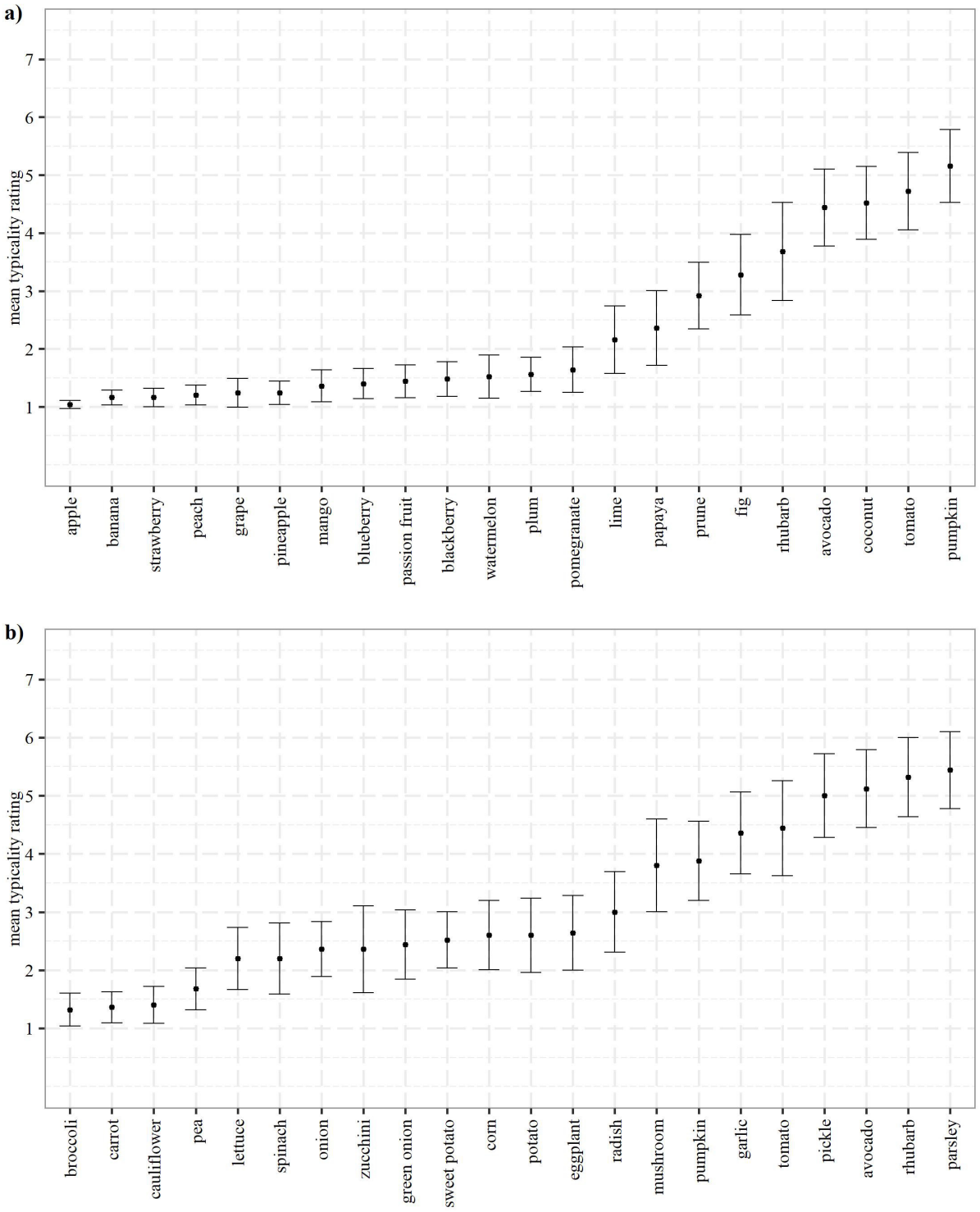

Figure 42: Mean and 95% confidence interval as error bars of the cleaned typicality ratings for a) fruit and b) vegetable.

The frequency distribution of the different ratings (Table 37) shows that the highest rating was chosen by far most for both categories: they constitute 50% of all ratings for *fruit* and 38% for *vegetable*. 5 and 6 were used the least often, for *fruit* each in only ~5% of the cases and for *vegetable* in ~7% of the cases.

Table 37: Frequency distribution of each typicality rating per category from experiment 3.

category		1	2	3	4	5	6	7
fruit	#	327	74	70	63	38	29	59
	%	49.5%	11.2%	10.6%	9.5%	5.8%	4.4%	8.9%
vegetables	#	266	93	67	83	52	47	96
	%	37.8%	13.2%	9.5%	11.8%	7.4%	6.7%	13.6%

The rating distributions for each SC are visualised in histograms in Figure 43 for *fruit* and Figure 44 for *vegetable*. They reveal that high means (<3) are a result of 1 being the dominant response and that means on the medium typicality level often are the result of 1 still being one of the dominant responses, but other ratings occurring with a higher frequency. Some of the medium ratings are multimodal. *Rhubarb*, for example, has a mean typicality of 3.8 for *fruit*, but it received 40% of 1 and 2 ratings, 33% of 6 and 7 ratings and only 27% between 3 and 5, where the mean lies. Its dip p-value is .00 and its IQR is 5. 27% of typicality rating distributions for *fruit* SCs and 45% for *vegetable* SCs have a dip p-value ≤.05 and a high IQR ≥3, indicating multimodality. For *fruit*, this applies to all SCs with a typicality smaller or equal to 2.9 and for *vegetable* to all SCs with a typicality smaller or equal to 2.5 with the exception of *rhubarb, avocado* and *pumpkin*, which have high IQRs, but dip p-values greater than .05. *Avocado* and *rhubarb* have unimodal distributions with a high skew, while the distribution for *pumpkin* is almost uniform.

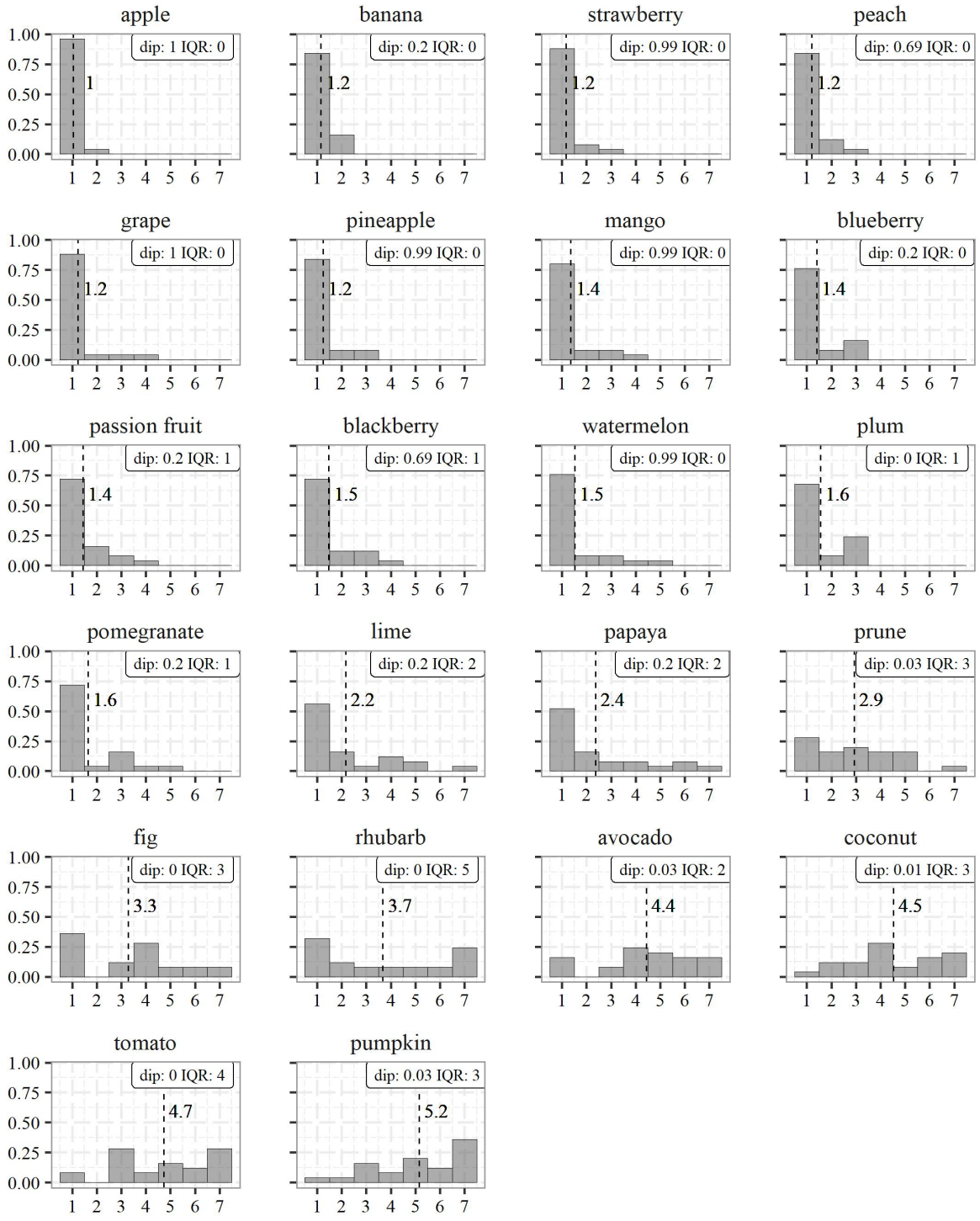

Figure 43: Distribution of fruit typicality ratings per SC in experiment 3 after cleaning with dip p-values (dip) and IQRs.

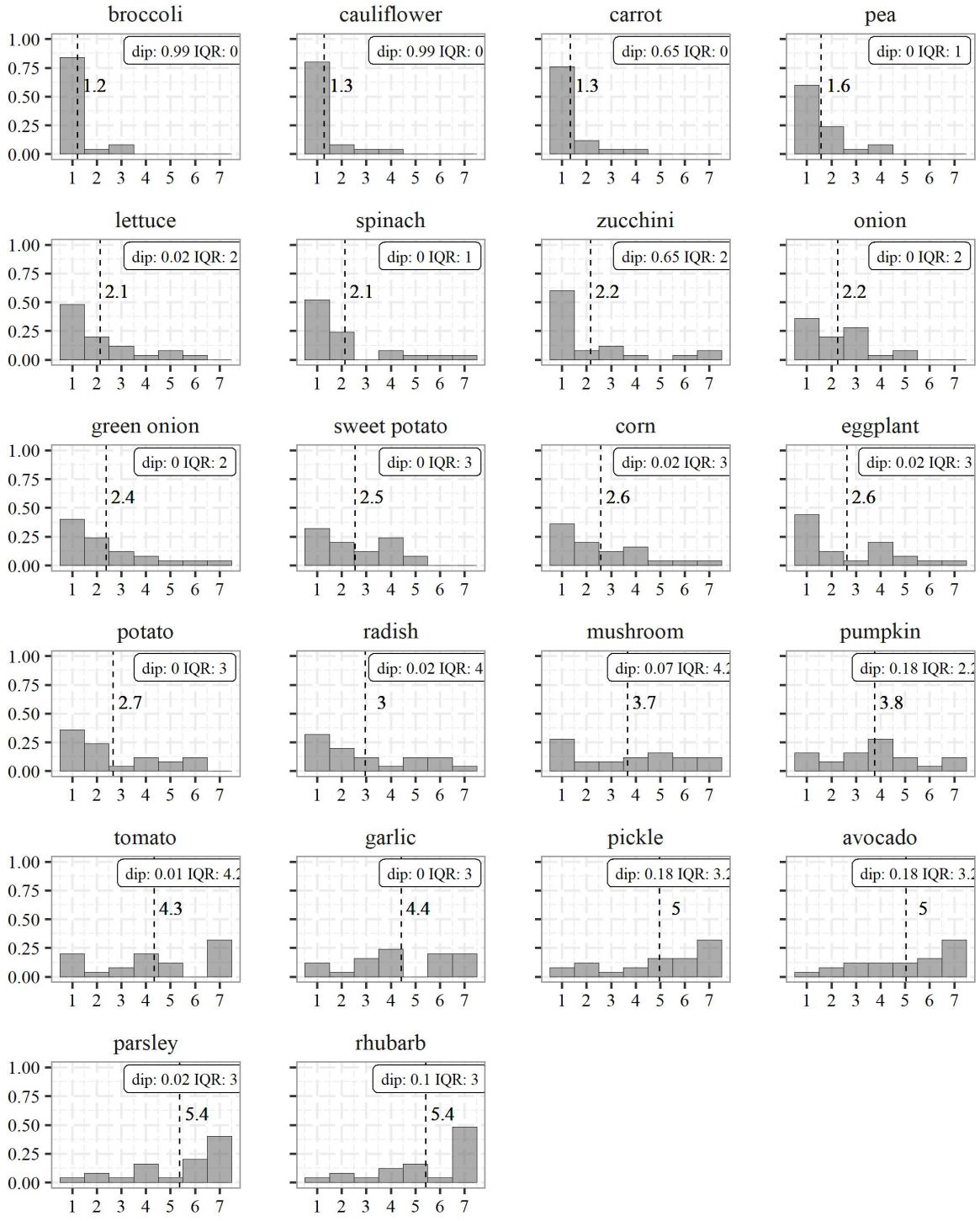

Figure 44: Distribution of vegetable typicality ratings per SC in experiment 3 after cleaning with dip p-values (dip) and IQRs.

Figure 45 shows the modes against the means for both categories. 1 is by far the most frequent mode. There are no SCs with a mode on 2, 5 and 6 and for *vegetable* additionally no modes on 3.

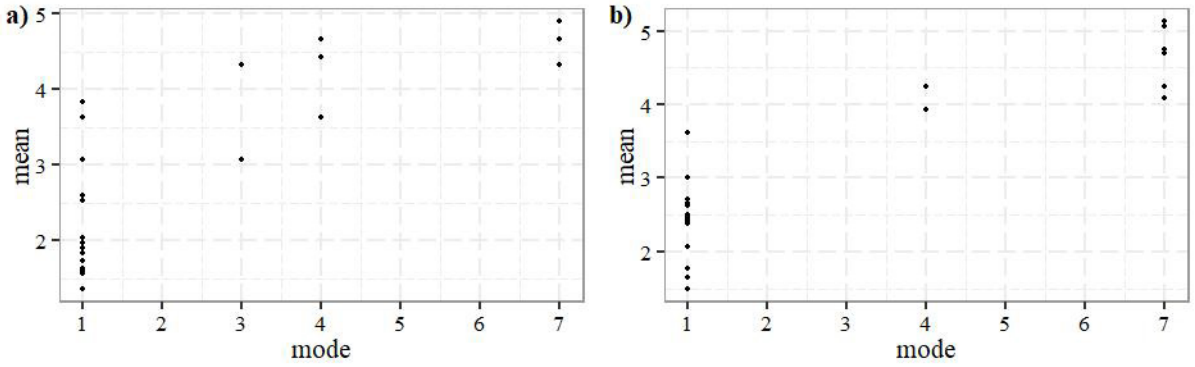

Figure 45: Mode(s) against mean for a) fruit and b) vegetable typicality ratings.

6.3.4.3 Discussion

The internal reliability of the ratings is high after exclusion of those participants with a low person-group-correlation. This procedure seems to be justified by the fact that the observed SDs were much higher compared to those found in the literature. It is possible that some participants placed random judgements to complete the task and get the reward faster. Excluding participants reduced the SDs to levels more in accordance with those reported in the literature but did not change the means substantially. The external reliability is very high for *fruit* with a mean correlation of .82 with other studies and high for *vegetable* with a mean correlation of .65.

The rating behaviour in experiment 3 matches the one of the other three studies for which rating data are available (see section 3.2.7). The highest rating was used disproportionally often and on the medium typicality level the ratings have high SDs and multimodal distributions.

The goal to use SCs that represent the whole typicality scale in equal proportions was not completely met. For *fruit*, the high typicality level is overrepresented, and the medium level is underrepresented. No SCs have a mode on 7 and are thus very untypical. For *vegetable*, the high typicality level is underrepresented, and the medium level is overrepresented. There are 6 SCs with mode 7.

6.3.5 Results: Probability ratings

The mean completion time for the probability ratings was 18:29 minutes (SD = 5:39 minutes), after exclusion of one outlier who took 40 minutes for the fruit questionnaire, and 21:36 minutes (SD = 15:32 minutes) for vegetable, after the exclusion of 2 outliers who took more than 30 minutes. The mean rating time of all properties for one SC was 52 seconds (SD = 19 seconds) for the fruit and 1:01 minutes (SD = 42 seconds) for the vegetable questionnaire. The participants took in general time to read the instructions and to complete the example with a mean of 1:18 minutes (SD = 29 seconds) for fruit and 1:13 (SD = 47 seconds) for vegetables.

6.3.5.1 Reliability

The person-group correlations of the property probability ratings for SCs are in Table 38. Their mean is .71 (SD = 0.09) for *fruit* and .68 (SD = 0.08) for *vegetable*. The lowest correlation is .43 for fruit and .48 for vegetable. Therefore, no need to exclude participants was seen.

Table 38: Person-group-correlations for probability ratings from experiment 3 for a) fruit and b) vegetable.

a) participant	correlation with remainder
1	0.78
2	0.66
3	0.72
4	0.72
5	0.72
6	0.80
7	0.73
8	0.77
9	0.66
10	0.79
11	0.64
12	0.57
13	0.83
14	0.73
15	0.68
16	0.71
17	0.81
18	0.63
19	0.74
20	0.77
21	0.55
22	0.63
23	0.75
24	0.78
25	0.66
26	0.43
27	0.79
28	0.82
29	0.78
30	0.78
Mean	0.71
SD	0.09

b) participant	correlation with remainder
1	0.63
2	0.69
3	0.74
4	0.48
5	0.80
6	0.74
7	0.71
8	0.69
9	0.63
10	0.69
11	0.45
12	0.67
13	0.65
14	0.74
15	0.62
16	0.73
17	0.63
18	0.72
19	0.62
20	0.66
21	0.70
22	0.65
23	0.76
24	0.71
25	0.73
26	0.71
27	0.81
28	0.52
29	0.72
30	0.74
31	0.76
32	0.65
Mean	0.68
SD	0.08

Another way to assess the coherence of the probability ratings is to compare the ratings for the borderline cases *avocado*, *pumpkin, rhubarb* and *tomato* from both versions of the questionnaire. It is remarkable that the means and SDs are almost identical despite the different contexts in the two experiments. The correlations between the two settings are between 0.95 and 0.98. Many SDs are high and therefore the intersubjective agreement is low in many cases.

6.3.5.2 Ratings

In what follows, the ratings are compared first on the category level, to show how the probability ratings differ between the fruit and vegetable category, and then on the SC level.

The highest and the lowest rating, 0 and 100, were used together in ~30% of the cases for both categories and both property combinations and the intervals between 0 and 10 and between 90 and 100 were used in ~50% of the cases. The other 50% of ratings are relatively evenly distributed (between 4% and 9%) among the remaining 8 intervals.

Figure 46 shows the probability ratings pooled over SCs per category and property. All properties selected for *fruit* received low ratings for *vegetable* SCs and vice versa. Exceptions are *has a heart* which received almost uniquely 0 ratings under 25 and *tastes tart*, which has a uniform distribution for *vegetable*.

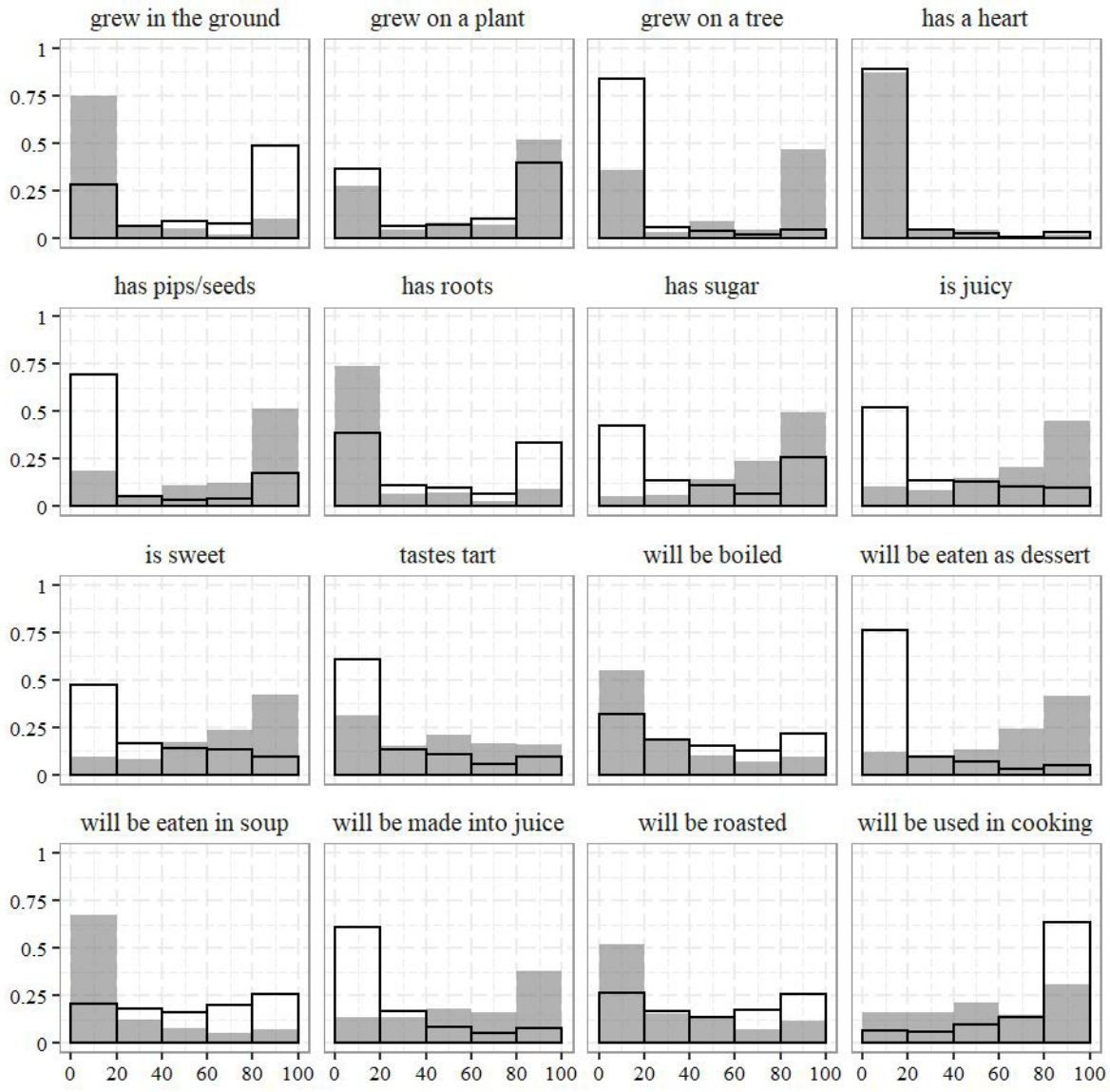

Figure 46: Histogram of probability ratings for fruit SCs (grey) and vegetable SCs (black lines) from experiment 3 per property.

Table 39 and Figure 47 show the frequencies of means, SDs and IQRs for fruit (top) and vegetables (bottom). Both *fruit* and *vegetable* SCs have approximately 35% of mean property

probability ratings between 50 and 75. For fruit, there are almost equally many between 75 and 100 and 50 and 74 and roughly 30% below 50. For vegetable, values under 50 dominate (55%) and roughly 20% are between 75 and 100. The SDs and IQRs are in general quite high. For both categories, only 20% of SDs are below 20 and most (45% for fruit, 40% for vegetables) are between 20 and 30. For *fruit*, 40% of the SDs are above 30 and for *vegetable* even 45%. For both categories, approximately 50% of IQRs are between 25 and 50, with 25% below 25 and 25% above 50. The means and SDs of all properties are in the appendix in Table 88 resp. Table 89.

Table 39: Distributions of means, SDs and IQRs for probability ratings from experiment 3 for fruit and vegetables for category properties.

means

category	[0,10)	[10,20)	[20,30)	[30,40)	[40,50)	[50,60)	[60,70)	[70,80)	[80,90)	[90,100]
fruit	4	2	12	12	24	19	35	43	37	10
	2.0%	1.0%	6.1%	6.1%	12.1%	9.6%	17.7%	21.7%	18.7%	5.1%
vegetable	16	11	17	15	12	21	26	20	14	2
	10.4%	7.1%	11.0%	9.7%	7.8%	13.6%	16.9%	13.0%	9.1%	1.3%

SDs

category	[0,10)	[10,20)	[20,30)	[30,40)	[40,50]
fruit	2	33	89	59	15
	1.0%	16.7%	44.9%	29.8%	7.6%
vegetable	7	19	61	57	10
	4.5%	12.3%	39.6%	37.0%	6.5%

IQRs

category	[0,10)	[10,20)	[20,30)	[30,40)	[40,50)	[50,60)	[60,70)	[70,80)	[80,90)	[90,100]
fruit	10	15	39	50	35	21	8	6	5	9
	5.1%	7.6%	19.7%	25.3%	17.7%	10.6%	4.0%	3.0%	2.5%	4.5%
vegetable	20	15	16	29	33	16	8	6	8	3
	13.0%	9.7%	10.4%	18.8%	21.4%	10.4%	5.2%	3.9%	5.2%	1.9%

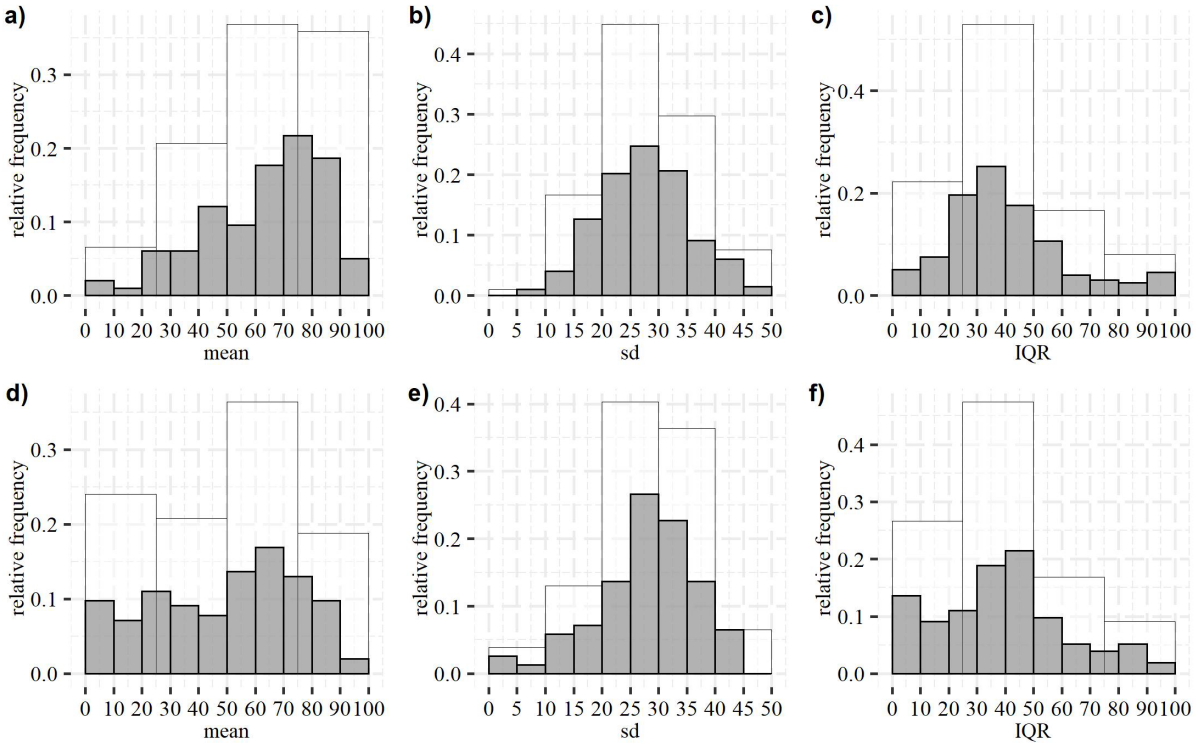

Figure 47: Relative frequencies of a), d) means, b), e) SDs and c), f) IQRs for fruit (top) and vegetable (bottom) SCs' property probability ratings from experiment 3 for category properties.

The high variance in the probability ratings led to a further inspection of the data. A high variance of mean property probability ratings shows that different individuals do not agree, and the mean thus cannot be assumed to reflect the common sense. It is not sensible to include them in the further analyses, which are supposed to reflect those parts of the mental representations of the categories that are shared between different individuals. The SD is very sensitive to outliers. For the further use of the data, it was therefore decided to base the exclusion of SC-property-pairs on the IQR. Figure 48 shows the IQRs plotted against the mean probability ratings. IQRs greater than 40 are exclusively found for properties with mean probability ratings in the medium probability range, between 25 and 75.

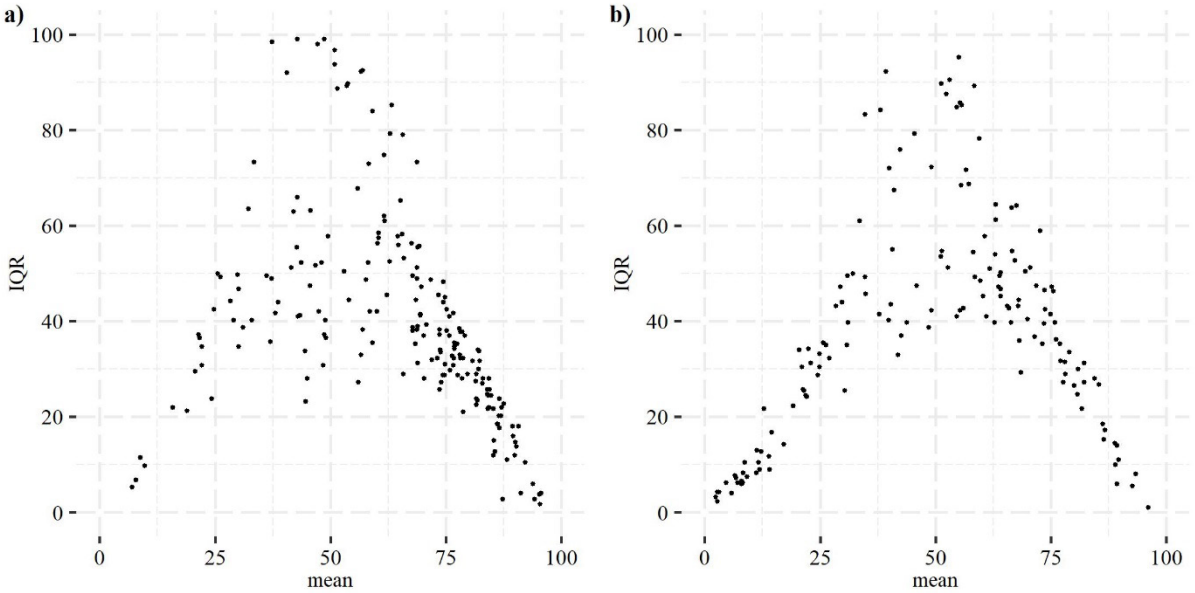

Figure 48: IQR against mean for property probability ratings for a) fruit and b) vegetables.

All IQRs per SC and property are in Table 40 for *fruit* and Table 41 for *vegetable* with the number of high IQRs (≥ 40) per property and SC. These tables show that there are both properties about which the participants did not agree as well as SCs. The total amount of SC-property-pairs with a high IQR is 84 (42%) for *fruit* and 73 (47%) for *vegetable*. For *fruit*, the properties with most high IQRs are *will be made into juice* (17 SCs), *grew on a plant* (16 SCs) and *tastes tart* (13 SCs). For *vegetable*, they are *has roots* (17 SCs), *will be eaten in soup* and *will be boiled* (14 SCs) and *grew in the ground* (13 SCs). The SCs with most high IQRs are *avocado* and *papaya* (7 properties) for *fruit*. For *vegetable*, they are *cauliflower, pumpkin, rhubarb* and *parsley* (5 properties).

Table 40: IQRs of all SC-property-pairs for fruit.

subcategory	grew on a plant	grew on a tree	has pips/seeds	has sugar	is juicy	is sweet	tastes tart	will be eaten as dessert	will be made into juice	IQR≥40
apple	98.5	4	6	21.8	28	28	36.5	32.3	33.8	1
banana	99	21	63.5	30	50	37.3	37.3	41.8	55.5	5
strawberry	1.8	5.3	32.3	18	24.5	16	49	20.3	58.5	2
peach	96.8	23.5	47.3	28	25.8	22.8	63	37	57.8	4
grape	12	49.8	38	18.5	22	25.8	35.8	32.3	28.8	1
pineapple	79.3	84	73.3	23.8	18	31.8	66	35.3	34	4
mango	89.3	15	73	27	29	21.8	40.3	22.5	45.5	4
blueberry	12	42.5	67.8	29.8	34.8	28.8	33	37.8	28	2
passion fruit	74.8	49.5	20.3	32.3	33	39.3	50.5	28.8	49	4
blackberry	12.8	23.8	51.3	31.8	37.8	41.5	35.5	34.3	58.3	3
watermelon	38.3	30.8	3.8	27.5	10.5	22	34.8	42.5	56	2
plum	92	17.8	37	23.8	31	30.8	42	32	57.5	3
pomegranate	93.8	48.3	13.8	35.5	41	27.3	51.8	35.3	41.3	5
lime	99	2.8	45	42	18.5	34.8	24.8	63.3	55.5	5
papaya	92.3	48.8	44.5	44	29	33.5	44	48.8	52.3	7
prune	62	55.8	53.3	32.8	30.8	52.5	52.3	31.3	37	5
fig	89.8	29	38.5	25.8	37.3	38	41	39	57.8	3
rhubarb	92.5	11.5	29.5	56.3	41.3	47.5	34	38.8	41.8	5
avocado	85.3	73.3	65.3	28	46.8	38.8	49.3	40.3	44.3	7
coconut	98	2.8	21.3	45.5	88.8	42	36.5	38.3	61	5
tomato	4	9.8	11	38.3	24.5	40.3	52.3	22	56.3	3
pumpkin	79	6.8	14.8	27.3	33.8	23.3	49.5	44.5	51.3	4
IQR≥40	16	8	10	4	5	5	13	6	17	84 42%

Table 41: IQRs of all SC-property-pairs for vegetables.

subcategory	grew in the ground	has a heart	has roots	will be boiled	will be eaten in soup	will be roasted	will be used in cooking	IQR≥40
broccoli	85.3	6	72.3	39.5	57.8	54.8	17.3	4
carrot	1	2.3	46.5	36.3	35.3	43.3	18.5	2
cauliflower	51.3	9	68.8	40.5	41	49.3	15.3	5
pea	33.3	6.3	31.3	36.8	54	47.3	26.8	2
lettuce	52.8	61	87.5	16.8	24.3	14.3	79.3	4
spinach	84.8	6.3	84.3	51.3	48.5	30.5	26.5	4
onion	10	8.3	59	33	50.5	49.5	14.5	3
zucchini	89.8	6.3	50	43.5	38.8	42.8	35.3	4
green onion	30	9	39.8	49.5	61.3	47.5	24.8	3
sweet potato	6	3.3	63.8	45.3	39.8	27.3	14	2
corn	92.3	10.5	43.3	39.8	41	42.8	28	4
potato	5.5	7.3	64.5	29	43.3	31.8	11	2
eggplant	67.5	8.3	72	37	39.8	29.3	27.3	2
radish	31.5	10.5	42.5	34	41.5	45.8	68.5	4
mushroom	46.3	4.3	83.3	35	54.8	54.5	21.8	4
pumpkin	85.8	7.8	76	42.3	44.5	50.3	33.5	5
garlic	47.3	7.5	89.3	28.8	51	36	8	3
tomato	39.8	6.5	35.5	40.3	45.3	47.3	31.3	3
pickle	44	4.3	12.8	25.5	22.3	25.8	53.5	2
avocado	34.3	11.8	24.5	21.8	32.3	25.5	46.8	1
rhubarb	64.3	13	78.3	71.8	35	55	47.5	5
parsley	95.3	4	90.5	30.5	42.3	49.3	41.5	5
IQR≥40	13	1	17	8	14	14	6	73 47%

There are 704 ratings distributions in total. To investigate the reason for the high variance in the data, the rating distributions of all SC-property-pairs were examined, grouping the ratings into 5 intervals: improbable [0,20), rather improbable [20,40), medium probable [40,60), rather probable [60,80) and probable [80,100]. The ratings in all intervals were counted and inspected. Examples for rating distributions with this grouping are in Figure 49. A cleaning procedure was used to smooth the data and make the outlier-sensitive mean more representative of the dominant opinion. Ratings were excluded as outliers if there were less than 25% (≤8 ratings) in the interval(s) more than one interval away from the interval that contains the dominant number of ratings. For example, if *probable* was the dominant interval, ratings in the adjacent interval *rather probable* are included, but ratings in the intervals *medium probable*, *rather improbable* and *improbable* are excluded. This applies for example to *banana* and the property *has sugar*: 25 ratings are between 60 and 100, but 5 are lower than 60. Those 5 were excluded, which raises the mean from 82.1 to 90.2 and reduces the SD from 22.8 to 13.1.

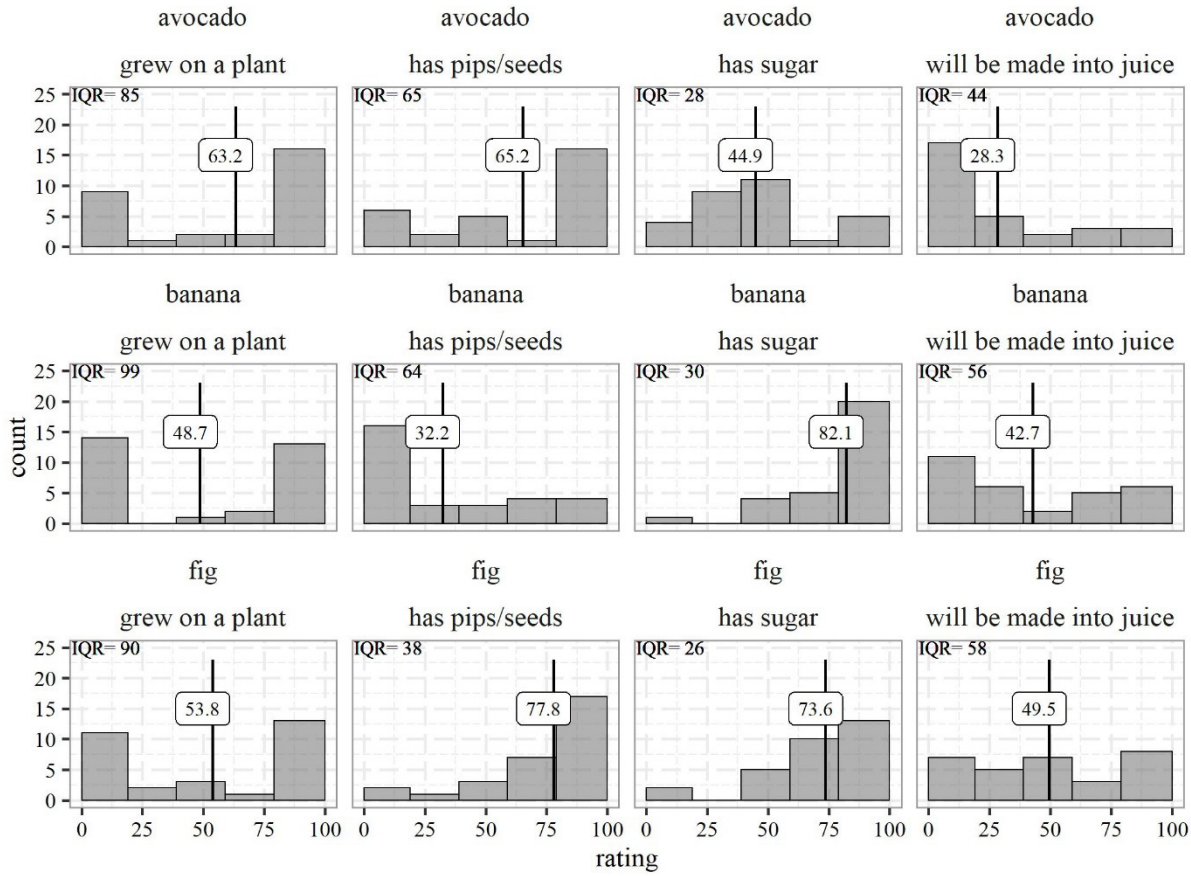

Figure 49: Rating distributions for the property ratings for grew on a plant, has pips/seeds, has sugar and will be made into juice for avocado, banana and fig with means (white box) and IQR (upper left).

Table 42 and Table 43 show the percentage of ratings outside the dominant interval and its adjacent interval if they are greater than 25%. Of the 198 fruit property probability rating distributions, 105 have less than 25% of ratings deviating from the dominant and dominant-adjacent interval. Of those, 4 had no deviating ratings at all and for 101 they were excluded. Only 8% (489) of the ratings were excluded. Of the 154 vegetable properties, 75 have less than 25% of ratings deviating from the dominant and dominant-adjacent interval. 5 had no deviating ratings at all and for 70 they were excluded. Only 6.5% (323) of the ratings were excluded.

Table 42: Percentage of ratings outside the dominant interval and its adjacent interval ≥25% for fruit.

subcategory	grew on a plant	grew on a tree	has pips/seeds	has sugar	is juicy	is sweet	tastes tart	will be eaten as dessert	will be made into juice	>25%	≤25%
apple	40 %						33 %			2	7
banana	53 %		37 %						43 %	3	6
strawberry							43 %		47 %	2	7
pineapple	43 %	47 %	37 %				53 %			4	5
grape		33 %	33 %							2	7
peach	57 %		30 %				47 %		50 %	4	5
plum	43 %							30 %	47 %	3	6
blackberry			33 %			37 %	30 %		47 %	4	5
watermelon									47 %	1	8
blueberry			53 %							1	8
passion fruit	43 %	43 %				30 %	50 %		37 %	5	4
mango	53 %		50 %				37 %		33 %	4	5
lime	43 %			47 %		37 %		50 %	37 %	5	4
papaya	47 %	33 %	33 %	33 %	43 %		47 %	40 %	57 %	8	1
pomegranate	57 %				30 %		53 %		37 %	4	5
fig	53 %				30 %	40 %		33 %	63 %	5	4
coconut	47 %			43 %	60 %			33 %	37 %	5	4
prune	50 %	40 %	43 %		30 %	40 %	60 %	30 %	33 %	8	1
tomato				33 %		33 %	53 %		43 %	4	5
pumpkin	33 %				33 %		53 %	53 %	43 %	5	4
rhubarb	47 %			47 %	33 %	57 %		30 %	40 %	6	3
avocado	40 %	33 %	43 %	30 %	37 %	33 %	33 %	33 %		8	1
>25%	16	6	10	6	8	8	13	9	17	93	
≤25%	6	16	12	16	14	14	9	13	5	105	

Table 43: Percentage of ratings outside the dominant interval and its adjacent interval ≥25% for vegetables.

subcategory	grew in the ground	has a heart	has roots	will be boiled	will be eaten in soup	will be roasted	will be used in cooking	>25%	≤25%
broccoli	53 %		53 %	28 %	41 %	34 %		5	2
carrot			28 %			41 %		2	5
cauliflower	38 %		50 %	31 %		47 %		4	3
pea				28 %		31 %		2	5
lettuce	34 %	34 %	59 %				53 %	4	3
spinach	50 %		34 %	44 %	47 %			4	3
onion			31 %		34 %	41 %		3	4
zucchini	59 %		38 %	44 %	34 %	41 %		5	2
green onion				31 %	38 %	31 %		3	4
sweet potato			38 %	50 %				2	5
corn	41 %		28 %	38 %	47 %	28 %		5	2
potato			44 %		31 %			2	5
eggplant	44 %		47 %	47 %	50 %			4	3
radish			31 %		50 %	44 %	44 %	4	3
mushroom	28 %		34 %	31 %	44 %	50 %		5	2
pumpkin	50 %		44 %	34 %	31 %	38 %		5	2
garlic	28 %		47 %		44 %	31 %		4	3
tomato				47 %	44 %	47 %		3	4
pickle	31 %						41 %	2	5
avocado							38 %	1	6
rhubarb	31 %		50 %	56 %		47 %	28 %	5	2
parsley	47 %		59 %		31 %	41 %	31 %	5	2
>25%	13	1	17	13	14	15	6	79	
≤25%	9	21	5	9	8	7	16	75	

This cleaning reduced the variability and the amount of IQRs over 40 from 84 to 74 for fruit and from 73 to 71 for vegetable. Furthermore, it reduced the mean IQR from 40.3 to 33 for fruit and from 38.9 to 35.3 for vegetable. Figure 50 shows the same SC-property-pairs as Figure 49 with ratings excluded in cleaning as white bars.

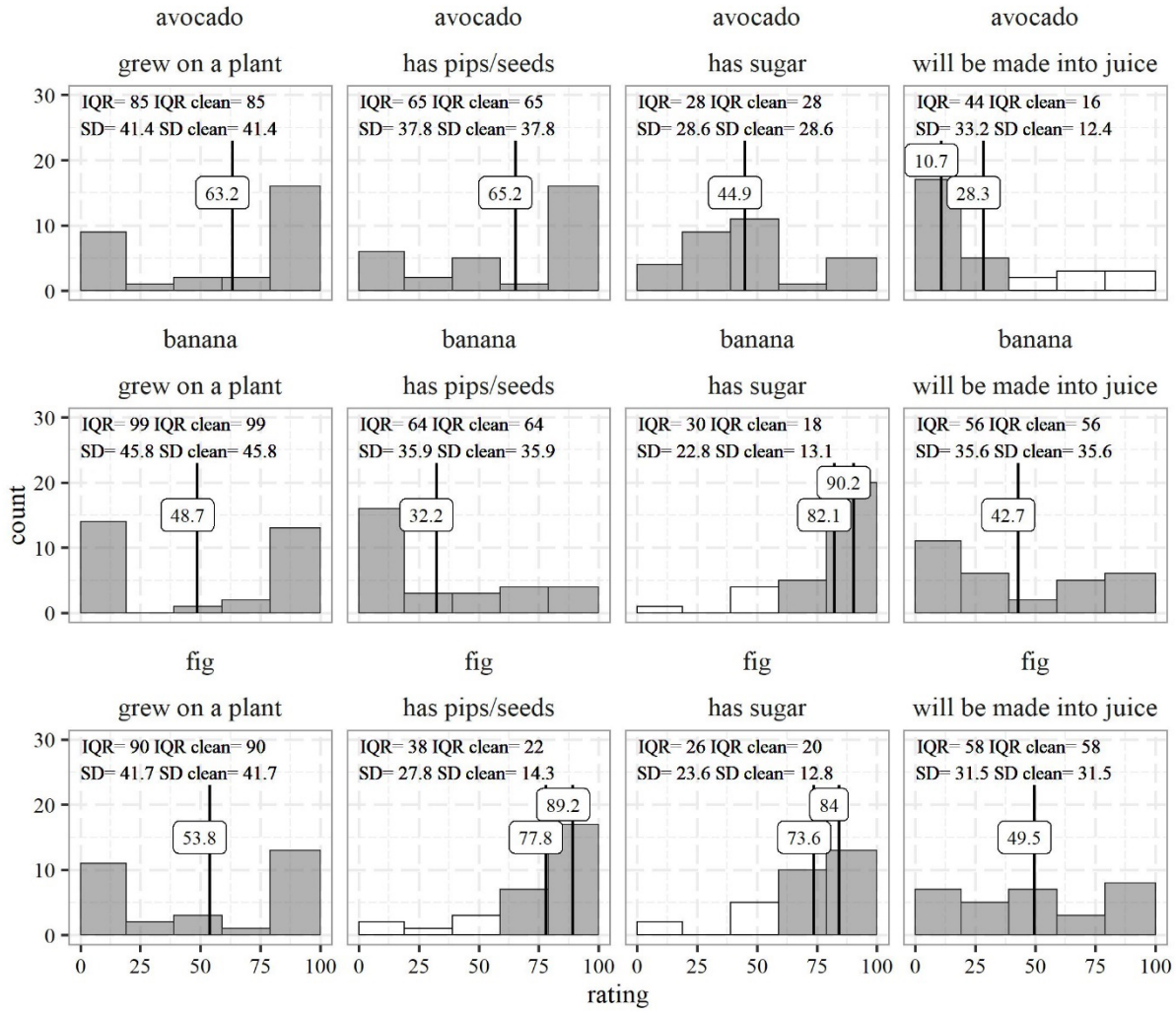

Figure 50: Rating distributions from figure 46 after cleaning with cleaned ratings(white bars), means before (lower) and after (higher) cleaning, as well as IQRs and SDs before and after cleaning.

Figure 51 compares the distributions of means, SDs and IQRs before and after cleaning for fruit and vegetables. For *fruit*, the amount of means between 80 and 100 is higher, while the amount of means between 70 and 80 is lower. More than 50% of the SDs are under 15 after cleaning, compared to under 10% before. The IQRs lie dominantly under 30 after cleaning, compared to under 40 before cleaning. For *vegetable*, there are 10% of means over 100 after cleaning, compared to under 5% before. The amount of means between 0 and 10 is 20% after cleaning, compared to 10% before. Almost 50% of the SDs are below 15 after cleaning, compared to 10% before cleaning. After cleaning, most IQRs are under 25.

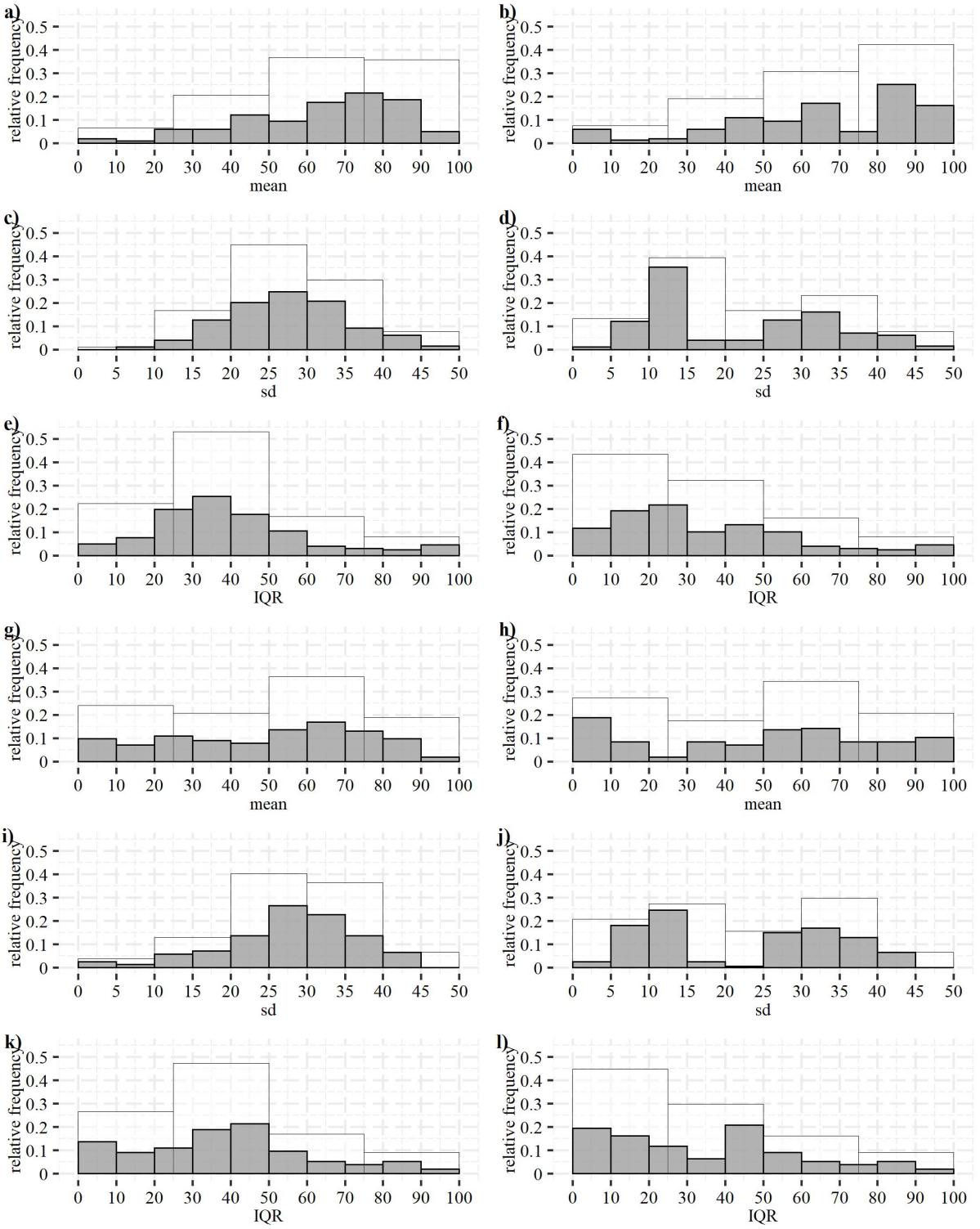

Figure 51: Comparison of distribution of means before (left) and after cleaning (right) for a), b) fruit and g), h) vegetables, as well as SDs before and after cleaning for c), d) fruit and i), j) vegetables and IQRs before and after cleaning for e), f) fruit and k), l) vegetables.

74 items (37%) for fruit and 71 items (46%) for vegetables have a high variance after cleaning. Figure 52 and Figure 53 show the first 32 SC-property-pairs that have a high variance after cleaning for *fruit* resp. *vegetable*. They either have multimodal distributions or uniform distributions or are highly skewed. Data with high variability and multimodal data are not well-described by their mean. As our formula uses means, all predictions were calculated in a cleaned version excluding these items.

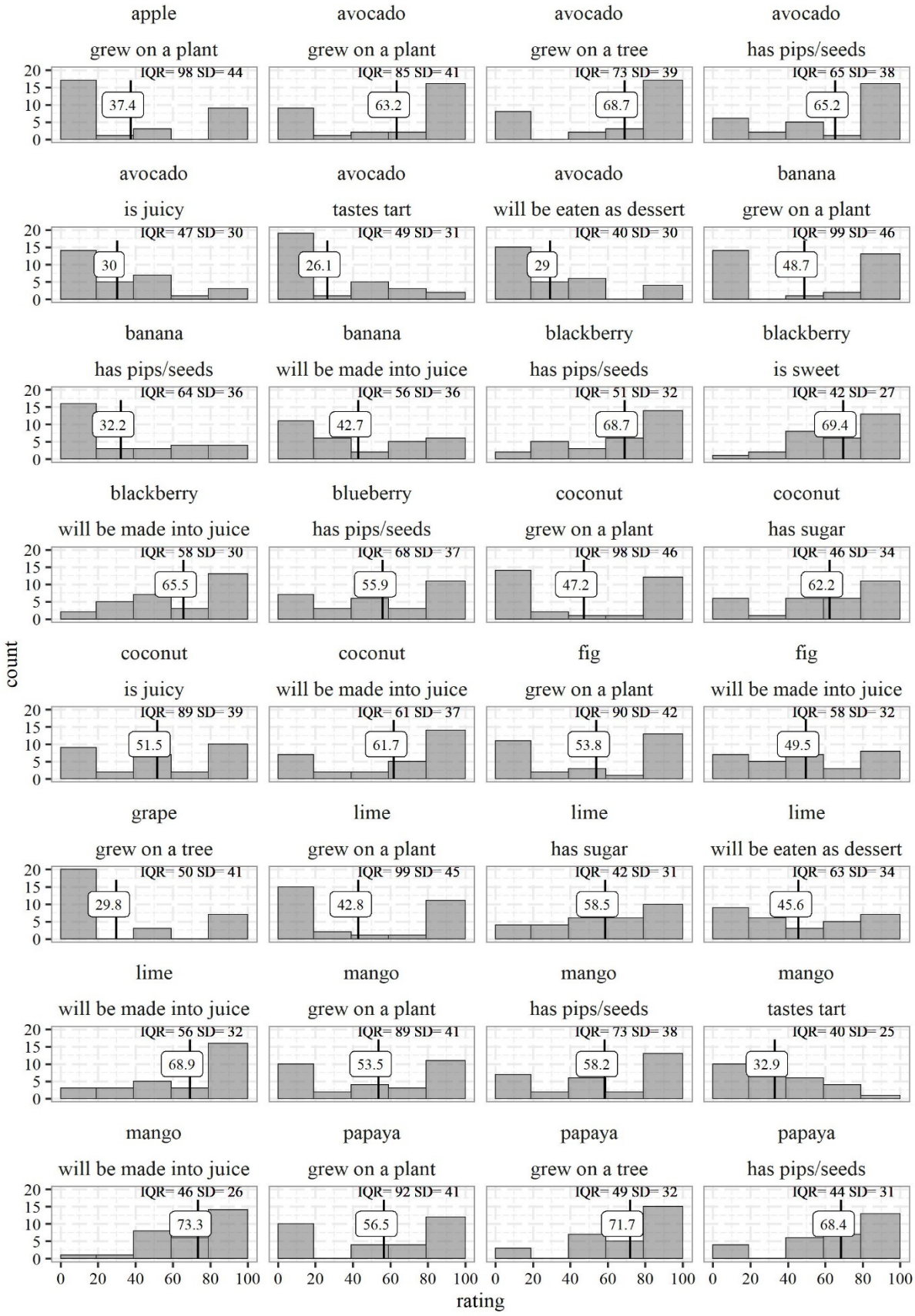

Figure 52: The first 32 SC-property-pairs with IQR >40 for fruit.

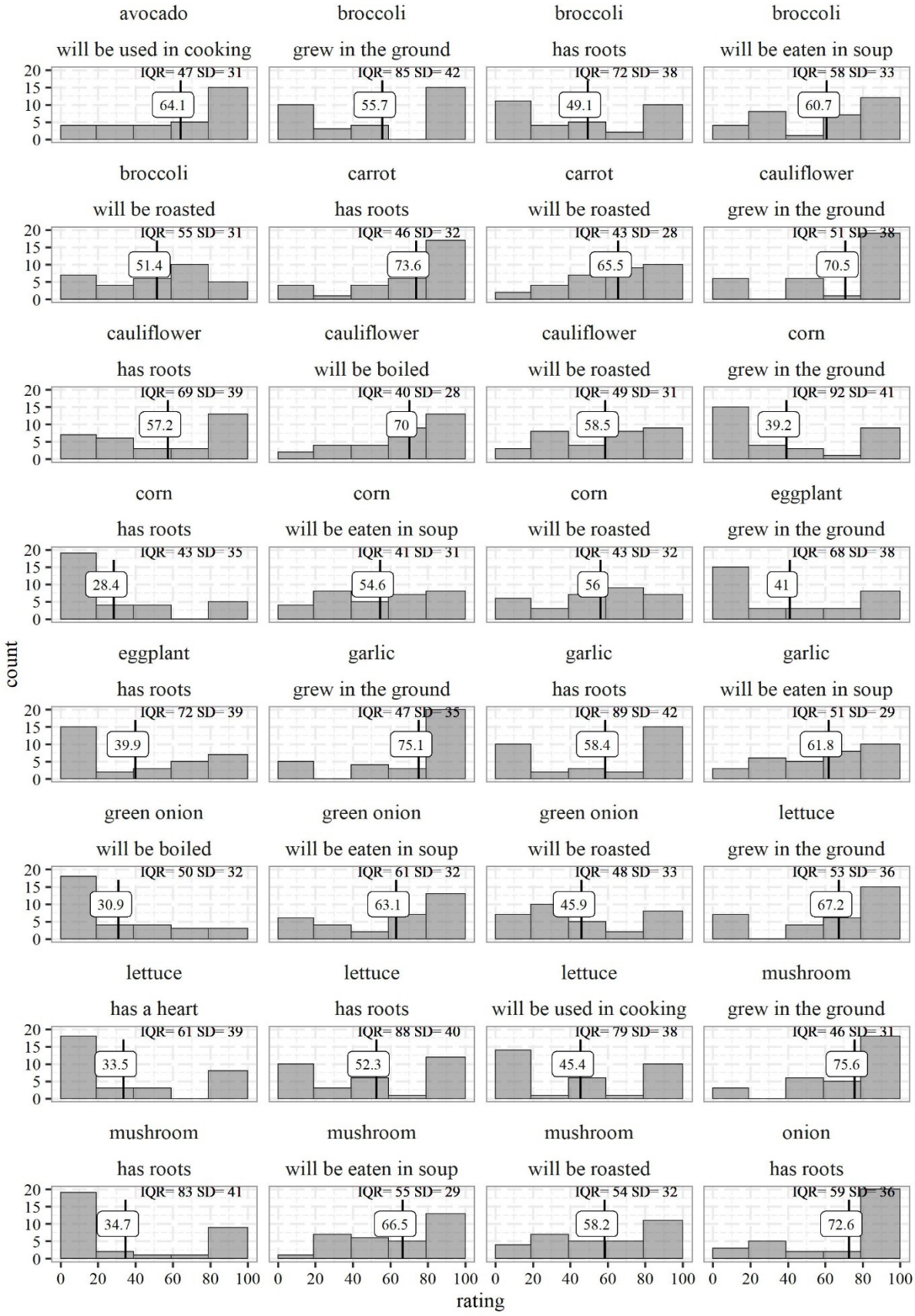

Figure 53: The first 32 SC-property-pairs with IQR>40 for vegetables.

6.3.5.3 Discussion

There is intersubjective stability for the probability ratings shown in high person-group-correlations and in the high correlation between the means and SDs of the borderline cases that were used in both experiments. Except for *has a heart*, which has a probability close to 0 for all SCs except one, the properties discriminate well between *fruit* and *vegetable*: fruit properties received a low probability rating for vegetable SCs and vice versa.

50% of the ratings are extreme ratings between 0 and 10 or between 90 and 100. While extreme ratings make sense for binary properties like growing conditions (SCs either grow on trees or not) and for properties that describe constituents (SCs either do or do not have seeds, roots, a heart and sugar), they are less plausible for preparation methods and common dishes. For these properties, there are usually different possibilities for each SC. For example, carrots are eaten in soups, but also in salads or raw. Participants who used extreme ratings in these cases presumably used probability ratings to express possibility and mean for example "it is 100% possible to eat carrots in soup" instead of "100% of carrots are eaten in soups". Similarly, medium probability ratings for binary properties express presumably uncertainty instead of probability and mean "I am 50% certain that avocados grow on trees" instead of "50% of avocados grow on trees".

The conflation of probability with possibility and uncertainty is presumably one reason for the high variances found. Another reason is the matrix design leading to SC-property-combinations that are untypical, for example people might have never thought about eating corn in soup. Probability ratings for them are probably not based on the stable cognitive representation of the SC and rather guessed ad-hoc. Some SC-property-pairs also could be interpreted in different ways, like *banana-has pips/seeds*. Bananas were bred to ensure that they do not contain mature seeds, but they do contain small, immature ones. Most subjects seem to agree that those should not be called seeds (most ratings lie between 0 and 10), but some seem to think they are seed-like enough to agree to this question, leading to some choosing higher ratings and some presumably expressing uncertainty ("it is possible to call them seeds but I would not") by choosing medium probability ratings.

Two strategies could be used to reduce the variance: binary properties seem to be better rated on 3-point-scale with the options "yes", "no" and "I don't know". Then, properties for which participants dominantly lack knowledge could be excluded. Avoiding the possibility reading could be done by pointing out more clearly that this is not what we ask for. Maybe by adding an example like

> Please note that we ask for the frequency of properties in the food items in question, not whether they are possible or to which degree they apply. A rating of 50 for the property tastes sweet means that you think that around 50% of these food items are sweet, not that they taste medium sweet compared to other objects. Likewise, a rating of 100 means that you think that all the objects taste sweet, not that it is possible for any one of them to taste sweet.

A last problem is the lack of alternatives for some properties. For example, cauliflower grows on the ground, but the only property included in the questionnaire was *grew in the ground*. The rating distribution has 3 modes, one each at very high, medium and very low probability, probably because some people thought "this is the closest description" and rated it highly probable, some people thought "this is half true" and rated it medium probable, and some thought "no, they grow *on* the ground" and rated it to be unlikely. The high dispersion which is presumably explained by a lack of alternatives might be avoided if the attributes to which the properties refer, like growing conditions, taste, typical dishes and preparation methods, would

be completed with more values. This might also help the participants to stick to the frequency reading. Our choice to include properties solely based on their ratings in the first two experiments could be improved by taking this into consideration.

6.4 General results and discussion

In this section, subjective probability ratings in general are analysed, making use of the questions from the introductory example in the questionnaires (section 6.4.1). Afterwards I investigate in how far these probability ratings can be said to follow the axioms of probability (section 6.4.2). Then, an application of Bayes' theorem on the data from experiment 1 and 2 is presented in section 6.4.3. Section 6.4.4 investigates the relationship between C and SC probability ratings, and section 6.4.5 the relationship between typicality and property probability ratings.

6.4.1 Subjective probability ratings

In this section, the answers to the questions from the introductory example about aquatic and land animals are discussed to gain some insights about the rating behaviour.

Experiment 1 asked participants to rate the diagnosticity, operationalised as Pr (C|P), of 8 properties in a contrastive way: they rated how probable it is that something is an aquatic animal and not a land animal based on the properties in question. The properties were intuitively chosen to include one that is discriminative for aquatic animals, *has gills*, two for land animals, *lives in trees* and *speaks English*, one that is not discriminating between the two because it applies to most animals, *has a liver*, and four that apply to both categories with different frequencies, *lays eggs, is yellow, swims* and *is small*.

The means, SDs and histograms for all ratings are in Figure 54. The two properties chosen for their discriminative power have the highest (has gills, mean = 4.7, SD = 1) and lowest (lives in trees, mean = -4.9, SD = 0.3). The third one, *speaks English*, which is clearly impossible for aquatic animals, received almost 50% of 0-ratings (aquatic and land animal equally probable).

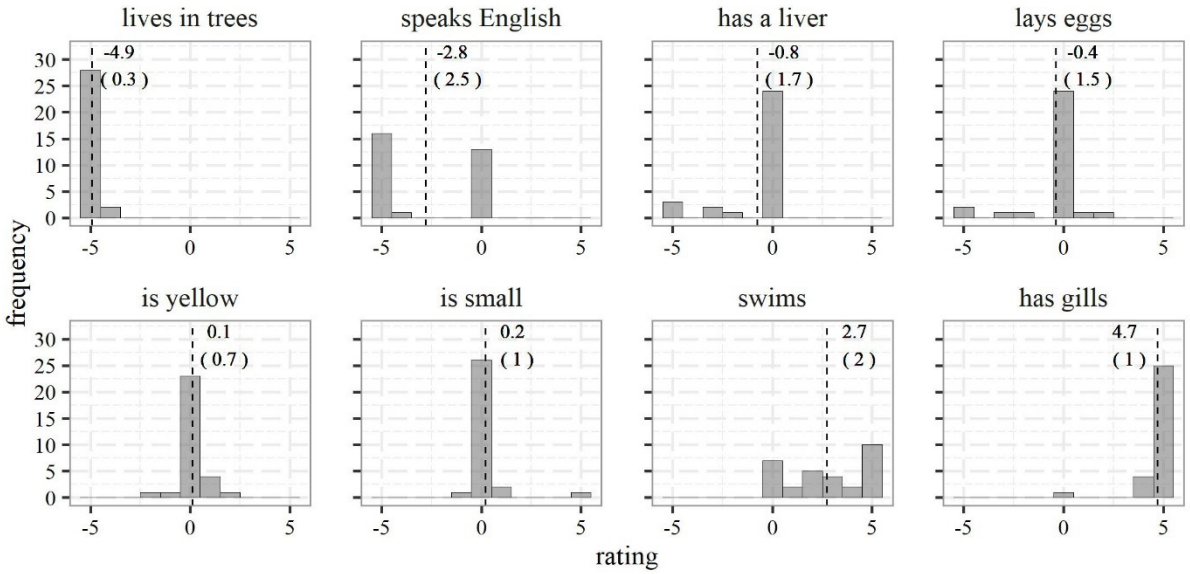

Figure 54: Histograms of ratings for example properties from experiment 1 with means (dashed lines) and SDs in brackets.

In experiment 2, 15 participants each rated the probability of the same properties, given they know that the animal in question is either aquatic or a land animal. Histograms with mean and

SD are in Figure 55. For both categories, there are 4 properties with an SD <1 and 4 that have a higher SD. The very probable (*swims* for aquatic animals) and very improbable (*speaks English* for aquatic animals, *has gills* for land animals) correctly only received the highest or lowest rating. For all other properties, the ratings have a clearly identifiable mode but higher SDs.

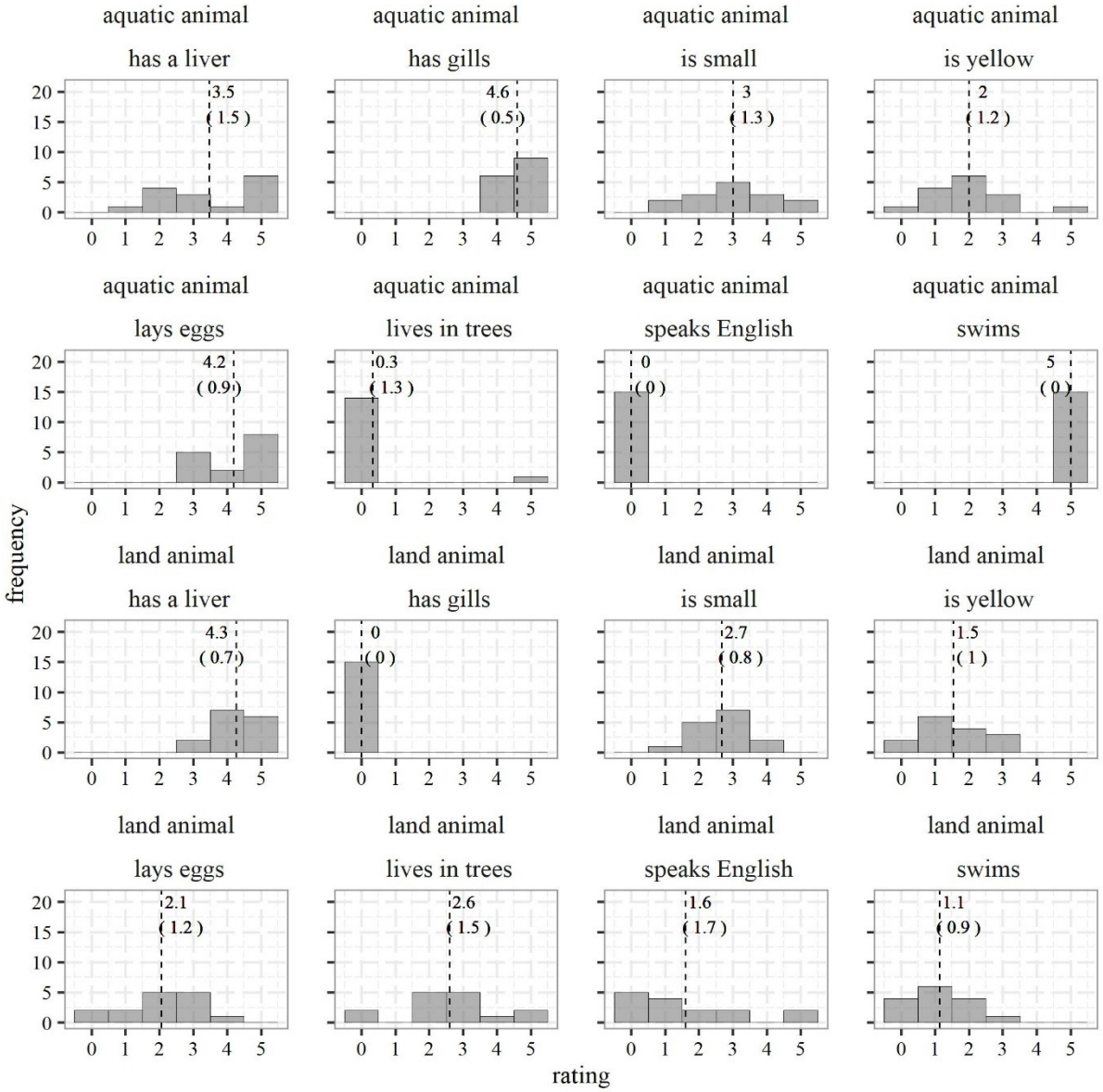

Figure 55: Histograms of ratings for example categories and properties from experiment 2 with means (dashed lines) and SDs in brackets.

In experiment 3, the same question as in experiment 2 for land animals was answered on a slider-scale from 0 to 100 by 62 participants. Histograms of the ratings with mean and SD are in Figure 56. The higher scale seems to introduce some uncertainty: the property *has gills* that is very improbable for land animals and that received uniformly 0 ratings in experiment 2, now received some ratings >10 and has a SD of 13. The property *is small* received some ratings greater than 80, while in experiment 1 the highest ratings were not used by anyone. The SDs are in general quite high: 6 properties, the rating distributions have a SD greater than 20.

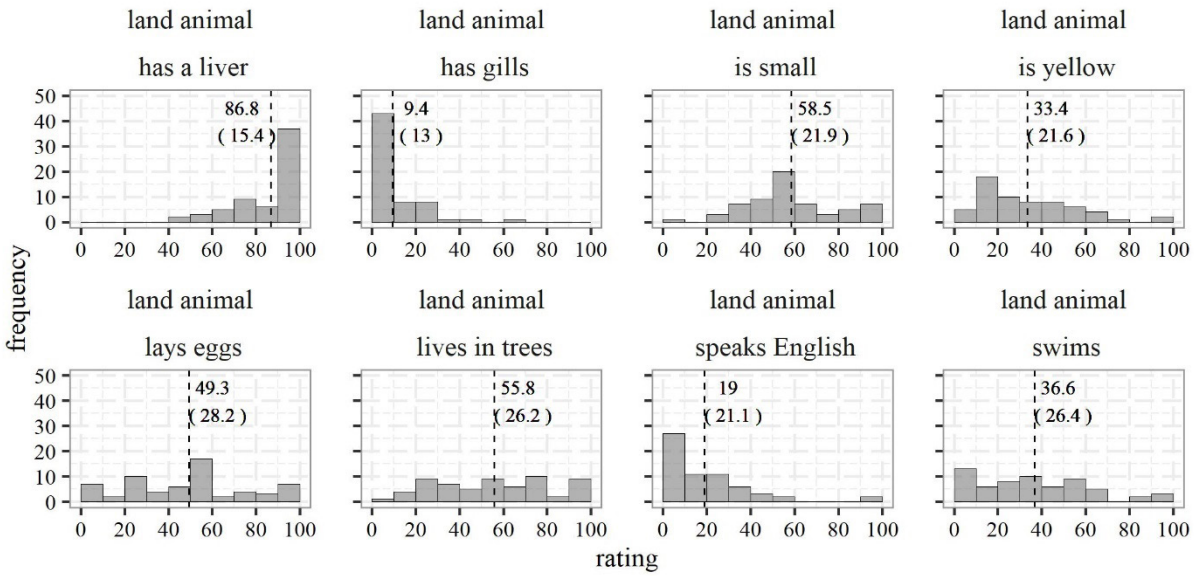

Figure 56: Histograms of ratings for land animal properties from experiment 3 with means (dashed lines) and SDs in brackets.

To compare the ratings of experiment 2 and 3, all ratings and means from experiment 3 were divided by 20 to fit on the same scale and plotted together in Figure 57. Despite the very unequal sample sizes – in the second experiment there were 15 and in the third 62 participants who answered the questions – the means have a low difference between 0.1 and 0.7. The highest differences are found for *speaks English* and *swims*, the former received more high ratings and the latter more 0 ratings in experiment 2.

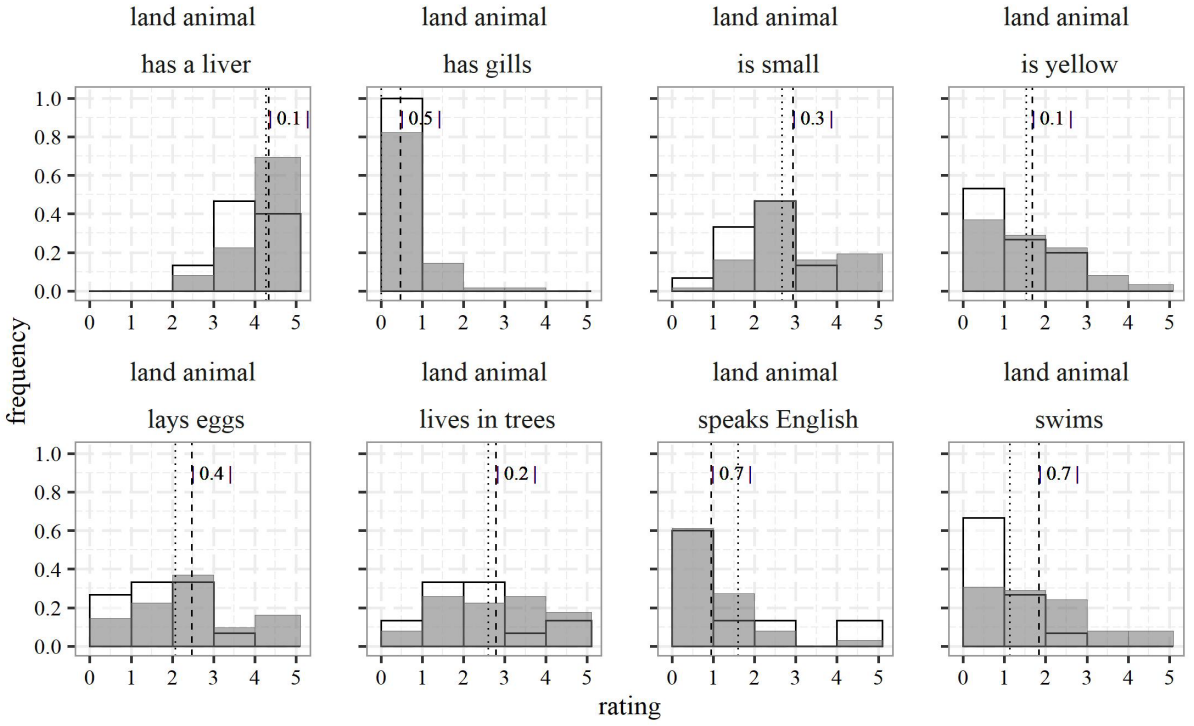

Figure 57: Ratings for land animal properties from experiment 2 (white bins) and experiment 3 divided by 20 (grey bins) compared with means (exp. 2 dotted, exp. 3 dashed) and the absolute difference between means.

This last comparison is a strong indication that probability ratings have intersubjective stability and that the means do not change substantially when more participants are included. The responses for medium probable properties have a high variance, but a mode exists in most cases.

6.4.2 Subjective probabilities and the laws of probability

The laws of probability or Kolmogorov axioms were presented in section 5.1. They are valid for all objective interpretations of probability and also assumed to be valid for epistemic probabilities for which it is usually assumed that they describe the grades of belief of rational agents which correspond or converge in most cases to the objective probabilities according to the statistical coordination principle. The subjective probabilities gathered in my experiments are different in that they are subjective estimations of objective frequencies which are objectively unknown and unknowable – it is impossible to know how many apples taste sweet or how many carrots are eaten in soup. All we assume is that for prototypical properties, the objective probabilities should correspond to the statistical majority and therefore be high.

Of the axioms and theorems presented in section 5.1, only one is reasonably testable for the probability data: upper boundary, i.e., 1 is the maximal probability. The maximal probability of exhaustive events $A_1, \dots, A_n$ is also 1:

$$\text{For each partition: } A_1, \dots, A_n: \sum_{1 \leq i \leq n} p(A_i) = 1 \text{ (Schurz, 2015, p. 11).}$$

The preparation methods (*will be boiled*, *will be roasted*) and common dishes (*will be made into juice*, *will be eaten as dessert*) from the experiments are not exhaustive but parts of the same partition and thus the sum of the probabilities of all properties describing them should be smaller than or equal to 1. For example, if I think that apples will be eaten as dessert in 80% of the cases, they cannot be made into juice in 40% of the cases. However, this all holds only if the properties are regarded as disjoint. This, however, is not always clear. For example, a participant might think that apples (not an individual apple but apples in general) can both be made into juice and eaten as dessert.

To test the disjointness of properties, I investigated whether this holds true for the mean values and for each participant individually. To do this, I counted the number of times in which the sum of the subjective probabilities of the properties that describe preparation methods for *vegetable* and the properties that describe common dishes for *fruit* is greater than 130. I added a slight tolerance and used 130 instead of 100 as a limit because it was difficult to set exact values and know which exact value was set with the slider scale. For the mean probabilities, 14 of the 22 SCs for *fruit* had a sum of probabilities greater than 130 and 3 SCs for *vegetable*. Figure 58 shows the percentage of inconsistencies with the law of upper boundary per participant. For *vegetable*, the sums of the probabilities are under 130 for 50% or less of the SCs for most participants. Only 4 participants had inconsistent ratings for more than 50% of the SCs. For *fruit*, one third of the participants rated inconsistently for over 75% of the SCs and half of the participants had less than 50% of inconsistent ratings.

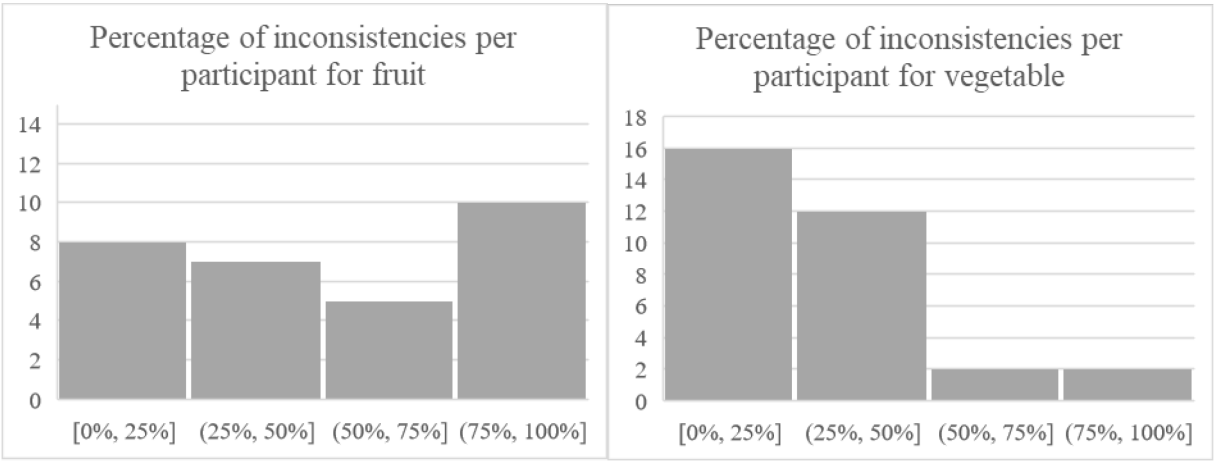

Figure 58: Percentage of inconsistencies with the probabilistic law of upper boundary per participant for fruit (left) and vegetable (right).

Ratings that are inconsistent with the law of upper boundary are frequently found for *fruit* and it is possible that they are less frequent for *vegetable* because the two preparation methods are not used at all for many of the SCs. This can mean two things: either the participants did not rate in a rational way or, as indicated in section 6.3.5.3, these participants rated the possibility of the properties for the SCs instead of their probability.

6.4.3 **Bayes' theorem applied**

Bayes' theorem allows to determine the conditional probability of C given P, $\mathrm{pr}(\mathrm{C}|\mathrm{P})$, if only the reverse probability of P given C, $\mathrm{pr}(\mathrm{P}|\mathrm{C})$, and the prior probabilities of A and B, $\mathrm{pr}(\mathrm{A})$ and $\mathrm{pr}(\mathrm{B})$, are known:

$$\Pr(\mathrm{A}|\mathrm{B}) = \frac{\Pr(\mathrm{P}|\mathrm{C}) \cdot \Pr(\mathrm{C})}{\Pr(\mathrm{P})} = \frac{\Pr(\mathrm{P}|\mathrm{C}) \cdot \Pr(\mathrm{C})}{\Pr(\mathrm{C}|\mathrm{P}) \cdot \Pr(\mathrm{C}) + \Pr(\mathrm{P}|\neg\mathrm{C}) \cdot \Pr(\neg\mathrm{C})}$$

His theorem derives directly from the definition of conditional probability, which states that the conditional probability of A given B is the probability of finding A and B together divided by the probability of finding B with and without A. In my experiments, I gathered probability estimations of categories C given properties P, $\Pr(\mathrm{C}|\mathrm{P})$, and of properties given categories, $\Pr(\mathrm{P}|\mathrm{C})$. If we assume that fruit and vegetable are clear contrast categories and that the priors are .5, i.e., that without additional information, it is equally likely that a food item is a fruit or vegetable, we can derive the following equality from Bayes' theorem for the subjective probabilities collected in the experiments (because the factor .5 cancels out):

$$\Pr(\mathrm{C}|\mathrm{P}) = \frac{\Pr(\mathrm{P}|\mathrm{C})}{\Pr(\mathrm{P}|\mathrm{C}) + \Pr(\mathrm{P}|\neg\mathrm{C})} \text{ or}$$

$$\Pr(\text{vegetable}|\mathrm{P}) = \frac{\Pr(\mathrm{P}|\text{vegetable})}{\Pr(\mathrm{P}|\text{vegetable}) + \Pr(\mathrm{P}|\text{fruit})}$$

The conditional probability on the left side of the equation were collected in experiment 1, and the probabilities on the right side were collected in experiment 2. As mentioned in section 5.1, it is not clear in how far subjectively estimated probabilities follow the laws for objective probabilities, because they tend to be biased for several reasons. Therefore, it is interesting to see in how far this theoretical equality holds for the empirical data. As there were different

participants in experiment 1 and 2, it is not possible to test conformity with Bayes' theorem on the participant level. But it is possible to test it for the means.

For example, the mean rating for *grows in the ground* is -3.27 in experiment 1, where -5 meant very probably a vegetable and 5 meant very probably a fruit, which amounts to a normalised .83 probability of a food item being a vegetable and not a fruit when it is known that it grows in the ground. In the second experiment, the normalised mean probability rating was .72 for *vegetable* and 0.13 for *fruit*. According to Bayes' theorem, the probability that something is a vegetable, given it grows in the ground, is with the data from experiment 2:

$$\Pr(\text{vegetable}|\text{grows in the ground}) = \frac{0.72}{0.72 + 0.13} = 0.85,$$

and, thus, almost identical to the normalised mean from experiment 1. This calculation was done for all 45 properties for which both kinds of ratings were available. A scatterplot of rated against calculated diagnosticity is in Figure 59, the correlation between the two is .97. The highest difference is for *is made of carbohydrate/starch* which received a .16 lower rating in experiment 1 (.25) than the one calculated from the ratings in experiment 2 (.41). The difference is ≥.1 in only two more cases: *is fried* and *grows on plants*. There is a remarkably high agreement of the calculated with the observed probabilities. This high correlation indicates that mean subjective probability estimations follow the laws of probability.

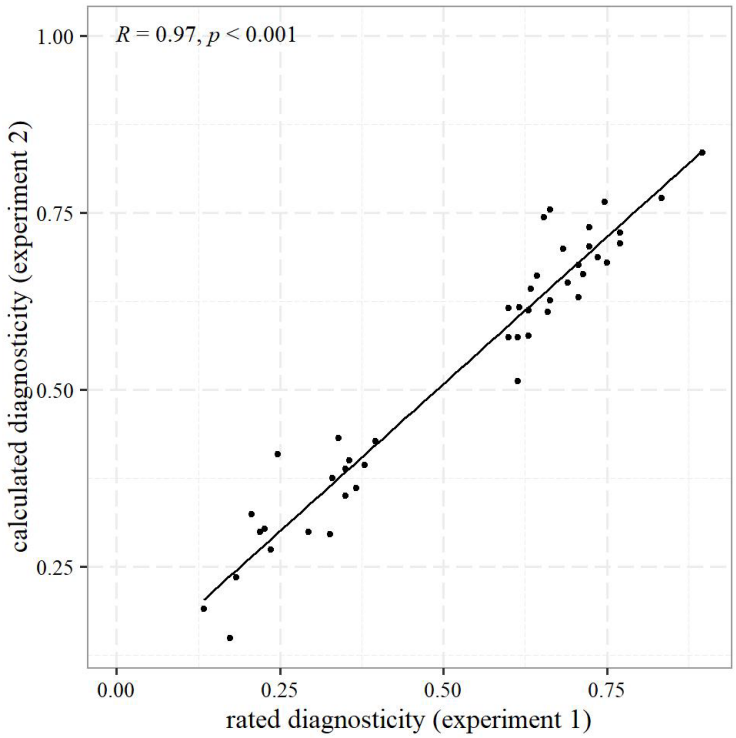

Figure 59: Scatterplot of rated diagnosticity from experiment 1 against calculated diagnosticity with the results from experiment 2.

6.4.4 Relationship between category and subcategory value probabilities

In this section, the relation between the property probabilities of Cs and SCs is discussed.

Table 44 shows the differences between the mean ratings from experiment 1 and 2 and the mean probability ratings from experiment 3 summarised over SCs for different typicality levels for *fruit*. The boundary for a high typicality is 1.4 and for medium typicality 3. The correlations of

the mean SC ratings from the high and medium typicality level with the data from experiment 2 are almost perfect with .94 and .97 and the one with the low typicality level is considerably lower with .67. The properties *will be made into juice*, *has pips/seeds*, *grew on a tree* and *tastes tart* have differences >.15 on all typicality levels. The correlations of the mean SC probability ratings with the diagnosticity ratings from experiment 1 are again very high for high and medium typicality levels with .95 and .96 and lower for the low level with .74.

Table 44: Ratings from experiment 1 and 2 and the difference to the means per typicality level from the SC ratings from experiment 3 with differences >.15 underlined for fruit.

		Pr(P|SC)					Pr(SC|P)			
property	Pr(P|C)	high	medium	low	all	Pr(C|P)	high	medium	low	all
will be made into juice	.93	<u>.24</u>	<u>.26</u>	<u>.46</u>	<u>.31</u>	.71	-.01	-.07	.08	-.02
has sugar	.92	.08	.12	<u>.33</u>	<u>.18</u>	.74	.09	.09	<u>.17</u>	.11
is sweet	.88	.06	.13	<u>.38</u>	<u>.20</u>	.77	.05	.08	<u>.18</u>	.10
is juicy	.88	.12	.13	<u>.38</u>	<u>.19</u>	.77	-.01	.02	<u>.19</u>	.07
has pips/seeds	.88	<u>.26</u>	<u>.18</u>	<u>.28</u>	<u>.21</u>	.71	-.15	-.09	.12	-.01
grew on a tree	.86	<u>.28</u>	<u>.26</u>	<u>.41</u>	<u>.30</u>	.72	-.08	-.11	.02	-.07
will be eaten as dessert	.84	.05	.11	<u>.33</u>	<u>.17</u>	.90	.05	.04	<u>.25</u>	.11
grew on a plant	.72	.07	.10	.09	.10	.61	.06	.08	.10	.09
will be used in cooking	.72	<u>.18</u>	<u>.21</u>	-.01	.15	.40	-.06	-.01	-.12	-.05
tastes tart	.59	<u>.20</u>	.14	<u>.19</u>	<u>.16</u>	.75	.03	.06	<u>.18</u>	.11
grew in the ground	.36	<u>.24</u>	<u>.21</u>	.05	<u>.17</u>	.23	.06	.05	-.15	-.01
has roots	.35	<u>.19</u>	<u>.18</u>	.07	<u>.15</u>	.22	-.02	-.01	<u>-.20</u>	-.07
will be boiled	.31	.07	.07	-.09	.02	.24	-.05	-.08	<u>-.32</u>	-.15
will be roasted	.25	-.01	.01	<u>-.22</u>	-.05	.18	<u>-.21</u>	-.13	<u>-.32</u>	<u>-.19</u>
will be eaten in soup	.21	.10	.07	<u>-.21</u>	.00	.13	-.03	-.06	<u>-.33</u>	-.15
has a heart	.17	.08	.08	.07	.08	.37	-.07	-.11	<u>-.23</u>	-.15
correlation		0.95	0.97	0.67	0.96		0.95	0.96	0.74	0.94

Table 45 shows the same data for *vegetable*. Here, the boundary for high typicality was chosen to be 2.2 and the boundary for medium typicality 3.5. The results are very similar with almost perfect correlations for the means of SCs on the high and medium typicality level and lower correlations for those on the low typicality level. The absolute values are however much lower for the SCs than for the category – many differences are >.15 on all typicality levels.

Table 45: Ratings from experiment 1 and 2 and the difference to the means per typicality level from the SC ratings from experiment 3 with differences >.15 underlined for vegetable.

		Pr(P|SC)					Pr(SC|P)			
property	Pr(P|C)	high	medium	low	all	Pr(C|P)	high	medium	low	all
will be used in cooking	.97	.29	.22	.29	.25	.60	.05	.01	.12	.05
will be eaten in soup	.88	.33	.32	.39	.35	.87	.03	.06	.33	.15
grew in the ground	.83	.23	.16	.31	.22	.77	-.06	-.05	.15	.01
will be roasted	.83	.38	.30	.35	.32	.82	.19	.13	.32	.19
has roots	.81	.35	.28	.43	.33	.78	.03	.01	.20	.07
will be boiled	.81	.24	.29	.49	.36	.76	.06	.08	.32	.15
grew on a plant	.69	.14	.15	.08	.13	.39	-.07	-.08	-.10	-.09
will be made into juice	.55	.28	.35	.26	.32	.29	.02	.07	-.08	.02
has pips/seeds	.45	.27	.27	.03	.19	.29	.07	.09	-.12	.01
has sugar	.42	-.02	-.02	-.03	-.03	.26	-.08	-.09	-.17	-.11
is sweet	.37	.05	.04	.02	.03	.23	-.05	-.08	-.18	-.10
is juicy	.34	.10	.08	-.02	.04	.23	-.01	-.02	-.19	-.07
grew on a tree	.32	.18	.20	.13	.18	.28	.08	.11	-.02	.07
has a heart	.31	.19	.21	.24	.22	.63	.08	.11	.23	.15
tastes tart	.28	.12	.08	-.03	.04	.25	-.04	-.06	-.18	-.11
will be eaten as dessert	.17	.02	.04	-.11	-.01	.10	-.05	-.04	-.25	-.11
correlation		0.91	0.90	0.69	0.87		0.96	0.96	0.74	0.94

This analysis shows that the estimated property probabilities for Cs and SCs are related and more so for the highly and medium typical SCs, as our formula would predict.

6.4.5 **Relationship between typicality und subcategory value probability**

This section analyses the relationship between typicality and property probability. One of our main assumptions is that the probability with which SCs have prototypical properties predicts their typicality rating (in a comparison with category probabilities, as shown in the formula presented in section 5.4). It is therefore interesting to analyse their relationship. The first question is in how far the probability of single properties predicts typicality and the second question is, if participants who judge typicality differently also judge property probability differently. If the latter is the case, this can be seen as a confirmation that typicality judgements are based on property probabilities.

Table 46 shows the correlation between the mean rated probability of a property given a subcategory and mean rated typicality of the subcategory. For *fruit*, typicality ratings are very highly correlated with the probability ratings for *has sugar* (.87), *will be eaten as dessert* (.78), *is sweet* (.73) and *will be eaten in soup* (-.72). For *vegetable*, the correlations are lower, but typicality ratings are correlated with the probability ratings for *will be boiled* (.57), *tastes tart* (-.57) and *will be eaten in soup* (.52).

Table 46: Correlation between each property's mean probability rating with mean typicality for fruit and vegetables.

		fruit		vegetable	
	property	r	p	r	p
fruit properties	has sugar	.87	.000	-.05	.836
	will be eaten as dessert	.78	.000	-.39	.075
	is sweet	.73	.000	-.02	.946
	will be made into juice	.59	.004	-.14	.528
	is juicy	.52	.013	-.20	.362
	grew on a tree	.19	.402	-.35	.105
	has pips/seeds	.15	.514	-.38	.080
	tastes tart	.08	.739	-.57	.005
	grew on a plant	.02	.922	-.04	.867
vegetable properties	will be boiled	-.44	.004	.57	.005
	will be eaten in soup	-.72	.013	.52	.013
	will be used in cooking	-.66	.054	.41	.057
	grew in the ground	-.43	.143	.30	.172
	has roots	-.47	.152	.29	.183
	will be roasted	-.58	.288	.22	.317
	has a heart	-.38	.625	.11	.623

Another interesting question is whether participants who rated typicality differently also rated the property probabilities differently. To identify relevant SCs, the percentage of ratings per scale point was examined and all SCs for which at least 5 participants (16.7% for fruit, 15.6% for vegetable) chose a rating different from the mode and not directly adjacent to the mode were examined. This criterion of multimodality applied to 8 fruit SCs and 13 vegetable SCs.

The probability ratings of the SCs with multimodal typicality ratings were examined. For this, the mean probability rating per property for each mode was calculated and the mean differences were compared in t-tests for unrelated data corrected for unequal sample sizes[31]. Out of 159 comparisons, 48 were significant with $p<.05$ for fruit and for vegetable 94 out of 336 comparisons were significant, both amounts to roughly 30% of significant differences between the modes. All SC-property-pairs with significant differences are in the appendix in Table 95 for fruit and Table 96 for vegetable SCs.

For example, participants who found it more likely that avocadoes taste tart (mean 61.5) also rated them to be untypical for fruit (typicality rating 7) than those who thought that it is improbable for avocadoes to taste tart (mean 14.3) and rated it to be medium typical (4). Participants who estimated coconuts to probably taste sweet (mean 77.1) also rated them to be more typical by choosing the rating 4 than those who rated them probably not sweet (mean 35.2) and chose the rating 7. Similar results were observed for *vegetable*. For example, participants who rated avocadoes to be unlikely juicy or consumed in juice also rated them to be atypical vegetables and participants who rated pumpkins less likely to be sweet also rated them to be more typical vegetables.

[31] $t = \frac{|\overline{x_a} - \overline{x_b}|}{\sqrt{\frac{\sum x_a^2 - \frac{(\sum x_a)^2}{N_a} + \sum x_b^2 - \frac{(\sum x_b)^2}{N_b}}{(N_a + N_b - 2)}} \left[\frac{N_a + N_b}{N_a N_b}\right]}$ (Coolican 2009, p. 361)

The fact that there were significant differences between the means of the probability ratings for participants who rated typicality differently shows that the reason for multimodal typicality rating distributions could be a difference in perception of the properties of the SCs. This is strong evidence for the assumption that typicality ratings are based on typical properties.

A similar question is investigated in Hampton&Passanisi (2016) and Djalal et al. (2018). While they found no significant correlations between participants' judgment of SC category membership and the properties that were generated for that category or attributed to that category, they found that the properties each single participant generated for or attributed to a category predicted their categorisation decisions.

7 Predicting typicality with probabilistic prototypes

This chapter presents the results from predicting typicality with the probabilistic prototype model. First with the means from the experiments (section 7.1) and then with individual ratings (section 7.2). Section 7.3 is a discussion of the results and section 7.4 presents some modifications that might improve the results.

7.1 Predictions with means

In chapter 5, I presented the probabilistic prototype model. This model predicts typicality from the diagnosticity-weighted min-similarity between the value probability distributions of SCs and Cs prototype. In a first step, the experimental data are used to construct frames for Cs and SCs (section 7.1.1). Then, the results of the typicality predictions that were calculated based on these frames are presented in section 7.1.2. Alternative models are developed to test the influence of the diagnosticity weights by leaving them out and applying them on the property instead of the attribute level (7.1.3). The last section 7.1.4 presents the results of parameter-fitting the diagnosticity weights to obtain optimal correlations.

7.1.1 Frames for Cs and SCs

The C prototype frames contain diagnosticity information for each attribute and probability information for each value. The SC frames contain probability information. With these two kinds of frames, typicality can be calculated as the diagnosticity-weighted average similarity with the formula

$$\mathrm{typ}(\mathrm{SC},\mathrm{C}) = \sum_{i=1}^{n} \mathrm{diag}(\mathrm{A_i}|\mathrm{C}) \cdot \mathrm{Sim}(\mathrm{SC},\mathrm{C}|\mathrm{A_i}),$$

in which similarity is defined as:

$$\mathrm{Sim}(\mathrm{SC},\mathrm{C}|\mathrm{A_i}) = \sum_{j=1}^{m_i} \min\left(\Pr\left(\mathrm{V_{ij}}\middle|\mathrm{C}\right), \Pr\left(\mathrm{V_{ij}}\middle|\mathrm{SC}\right)\right),$$

and diagnosticity as

$$\mathrm{diag}(\mathrm{A_i}|\mathrm{C}) = \frac{\max\left(\Pr(\mathrm{C}|\mathrm{V_{i1}}), \ldots, \left(\Pr\left(\mathrm{C}\middle|\mathrm{V_{im_i}}\right)\right)\right.}{\sum_{i=1}^{n} \max\left(\Pr(\mathrm{C}|\mathrm{V_{i1}}), \ldots, \Pr\left(\mathrm{C}|\mathrm{V_{im_i}}\right)\right)}.$$

How these frame components – attribute-value-assignment, diagnosticities and value probabilities – were derived from the experimental data will be discussed successively in what follows.

The attribute-value-assignment is in Table 47. It was generated based on the assignments of five members of our research group with the goal of finding mutually exclusive values, similar to the procedure described in Smith et al. (1988), but with five instead of two judges. The assignments of each researcher are in the appendix in Table 92. While the wording between researchers differed (e.g., for *eaten as dessert* there were CONSUMED-IN, USAGE, TYPICALLY-USED-IN and INGREDIENT-IN), there was general agreement on which properties have the same underlying attribute. For comparability, the wording from Smith et al. (1988) was used where possible. The attribute ROOTS from *has roots* does not appear in their list and their attribute HOW-EATEN was split into CONSUMED-IN (with values Dessert, Soup, Juice), PREPARATION-METHOD (with values Boiled and Roasted) and USED-IN-COOKING (with value Yes) because three out of five judges felt that those are important nuances.

Table 47: Attribute-value-assignment for the 16 properties used in model predictions.

property	new wording (ex. 3)	attribute	value
is eaten as dessert	will be eaten as dessert	CONSUMED-IN	Juice
is made into juice	will be made into juice	CONSUMED-IN	Dessert
is for soup	will be eaten in soup	CONSUMED-IN	Soup
has a heart		HEART	Yes
grows on trees	grew on trees	HOW-GROWING	Trees
grows on plants	grew on a plant	HOW-GROWING	On Plants
grows in the ground	grew in the ground	HOW-GROWING	In the Ground
has sugar		MAIN-NUTRITIONAL-COMPONENT	Sugar
is boiled	will be boiled	PREPARATION-METHOD	Boiled
is roasted	will be roasted	PREPARATION-METHOD	Roasted
has roots		ROOTS	Yes
has pips/seeds		SEEDS	Yes
is sweet		TASTE	Sweet
tastes tart		TASTE	Tart
is juicy		JUICINESS	Juicy
is used in cooking	will be used in cooking	USED-IN-COOKING	Yes

The value frequencies were computed from the means of the $\Pr(P|C)$ and $\Pr(P|SC)$ ratings gathered in experiments 2 and 3. For each attribute, the sum of all values must be 1 because this is the maximal probability. However, the sum of the mean normalised ratings is greater than 1 for some attributes (e.g., TASTE with $\Pr((\text{TASTE} = \text{Sweet})|\text{fruit}) = 0.88$ and $\Pr(\text{TASTE} = \text{Tart}|\text{fruit}) = 0.59$). They were normalised such that the value probabilities for each attribute have a sum of maximally 1 by normalising each value probability with the sum of value probabilities per attribute (e.g., $\Pr(\text{TASTE} = \text{Sweet}|\text{fruit}) = \frac{0.88}{0.88+0.59} = 0.6$). Some value probabilities have a sum smaller than 1 (e.g. PREPARATION-METHOD with $\Pr(\text{PREP.}-\text{METHOD} = \text{Roasted}|\text{fruit}) = 0.25$ and $\Pr(\text{PREP.}-\text{METHOD} = \text{Boiled}|\text{fruit}) = 0.31$) and for some attributes only one value probability was gathered (e.g. SEEDS with only $\Pr(\text{SEEDS} = \text{Yes}|\text{fruit}) = 0.88$). All attributes with a sum of value probabilities smaller than 1 were left as they were, because, although they could easily be completed by adding a value (Others or No), whose probability can be calculated with $\Pr(\neg P|C) = 1 - \Pr(P|C)$, their diagnosticity weight is not computable from the given data because the reverse is not true: $\Pr(C|\neg P) \neq 1 - \Pr(C|P)$.

The diagnosticity weights were calculated from the mean Pr(fruit|P) ratings[32] gathered in experiment 1. The probabilities for vegetables derive from these ratings when they are multiplied with -1, $\Pr(\text{vegetable}|\text{P}) = (-1) * \Pr(\text{fruit}|\text{P})$, as the lowest possible rating corresponds to "very probably a vegetable based on this property". The 11-point-scale used in experiment 1 was split into two scales from 0 to 5 and with $\Pr(\text{C}|\text{V}_{ij}) = \Pr(\text{C}|\text{P}_k)$ according to the attribute-value assignment above. The results were then divided by 5 to normalise between 0 and 1:

$$\Pr_{\text{norm}_{\text{category}}}(\text{fruit}|\text{V}_{ij}) = \begin{cases} 0, & \Pr(\text{fruit}|\text{P}) \leq 0 \\ \frac{\Pr(\text{fruit}|\text{P})}{5}, & \Pr(\text{fruit}|\text{P}) > 0 \end{cases}, \text{ and}$$

$$\Pr_{\text{norm}_{\text{category}}}(\text{vegetable}|\text{V}_{ij}) = \begin{cases} 0, & \Pr(\text{fruit}|\text{P}) \geq 0 \\ (-1) \cdot \frac{\Pr(\text{fruit}|\text{P})}{5}, & \Pr(\text{fruit}|\text{P}) < 0 \end{cases}.$$

The diagnosticity of an attribute A_i derives from these value diagnosticities as the maximum of A_i, divided by all attributes' maxima to normalise the typicality prediction to values between 0 and 1. As described above, in the chosen experimental setup the value probabilities do not add up to 1 for all attributes. To keep typicality predictions between 0 and 1, each maximum was divided by the sum product of the category value probabilities with the attribute weights:

$$\text{diag}(\text{A}_\text{i}|\text{C}) = \frac{\max\left(\Pr\left(\text{C}|\text{V}_{\text{i}1}\right), \ldots, \Pr\left(\text{C}|\text{V}_{\text{im}_\text{i}}\right)\right)}{\sum_{\text{i}=1}^{\text{n}} \max\left(\Pr\left(\text{C}|\text{V}_{\text{i}1}\right), \ldots, \Pr\left(\text{C}|\text{V}_{\text{im}_\text{i}}\right)\right)} \cdot \frac{1}{\sum_{\text{i}=1}^{\text{n}} \sum_{\text{j}=1}^{\text{m}_\text{i}} \Pr\left(\text{V}_{\text{ij}}|\text{C}\right)}.$$

The diagnosticity formula below differs from the mathematically regular diagnosticity formula on pages 89 and 149 in containing the additional factor $\frac{1}{\sum_{\text{i}=1}^{\text{n}} \sum_{\text{j}=1}^{\text{m}_\text{i}} \Pr\left(\text{V}_{\text{ij}}|\text{C}\right)}$. This is due to the sum of the subjectively estimated value probabilities, being sometimes smaller than 1, because only one attribute value was included, like for JUICINESS. This has the effect that the maximal similarity per attribute and therefore then also the maximal typicality of a subcategory is smaller than 1, if computed with the regular formula. To compensate for this problem stemming from the experimental design, the diagnosticities are additionally divided by the sum $\sum_{\text{j}=1}^{\text{m}_\text{i}} \Pr\left(\text{V}_{\text{ij}}|\text{C}\right)$. Using this corrected formula, the obtained typicality values are normalised between 0 and 1, as it should be. The results of each step of these calculations for category properties are in Table 48 for fruit and in Table 49 for vegetables. For the mean Pr(V|C) ratings there is an extra column entitled norm per A which reflects the normalisation of value probabilities such that they sum up to 1 per attribute for all attributes which had a sum greater than 1.

[32] In what follows, Pr(x) refers to the mean rated subjective probability of x from the experiments.

Table 48: Value probability and diagnosticity calculations for fruit with category properties.

property	attribute	value	mean Pr(V|C) rating	norm per A	mean Pr(C|P) rating	mean Pr(C|P) norm	max norm	diag A
will be made into juice	CONSUMED-IN	Juice	0.93	0.53	2.07	0.41	0.79	0.26
will be eaten as dessert	CONSUMED-IN	Dessert	0.84	0.47	3.97	0.79	0.79	0.26
grew on a tree	HOW-GROWN	Trees	0.86	0.54	2.23	0.45	0.45	0.15
grew on a plant	HOW-GROWN	On Plants	0.72	0.46	1.13	0.23	0.45	0.15
is juicy	JUICINESS	Juicy	0.88	0.88	2.70	0.54	0.54	0.18
has sugar	MAIN-NUTRITIONAL-COMPONENT	Sugar	0.92	0.92	2.37	0.47	0.47	0.15
has pips/seeds	SEEDS	Yes	0.88	0.88	2.13	0.43	0.43	0.14
is sweet	TASTE	Sweet	0.88	0.60	2.70	0.54	0.54	0.18
tastes tart	TASTE	Tart	0.59	0.40	2.50	0.50	0.54	0.18

Table 49: Value probability and diagnosticity calculations for vegetables with category properties.

property	attribute	value	mean Pr(V|C) rating	norm per A	mean Pr(C|P) rating	mean Pr(C|P) norm	max norm	diag A
will be eaten in soup	CONSUMED-IN	Soup	0.88	0.88	-3.67	0.73	0.73	0.30
has a heart	HEART	Yes	0.31	0.31	-1.33	0.27	0.27	0.11
grew in the ground	HOW-GROWN	In the Ground	0.83	0.83	-2.73	0.55	0.55	0.22
will be roasted	PREPARATION-METHOD	Roasted	0.83	0.50	-3.17	0.63	0.63	0.26
will be boiled	PREPARATION-METHOD	Boiled	0.81	0.50	-2.63	0.53	0.63	0.26
has roots	ROOTS	Yes	0.81	0.81	-2.80	0.56	0.56	0.23
will be used in cooking	USED-IN-COOKING	Yes	0.97	0.97	-1.03	0.21	0.21	0.08

The resulting probabilistic prototype frames are in Figure 60 for *fruit* and Figure 61 for *vegetable*. Both have six attributes. For *fruit*, three attributes have two values, and three attributes have one value, while for vegetables only one attribute has two values and the other five have one. The attribute CONSUMED-IN has the highest diagnosticity (0.26 resp. 0.3) for both categories. For a more complete graphic presentation, attributes with only one value were completed with Others or No values, but these values were not used in the calculations.

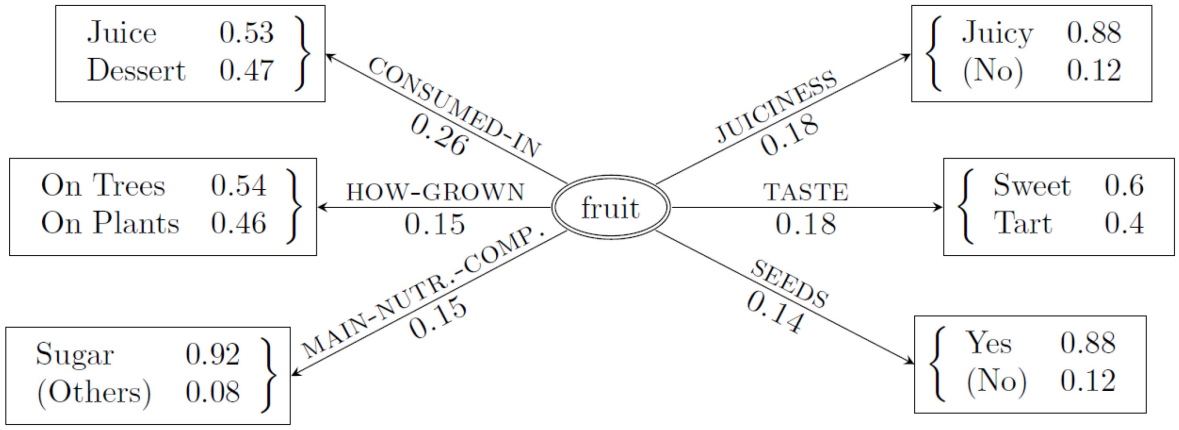

Figure 60: Prototype frame for fruit resulting from experiment data.

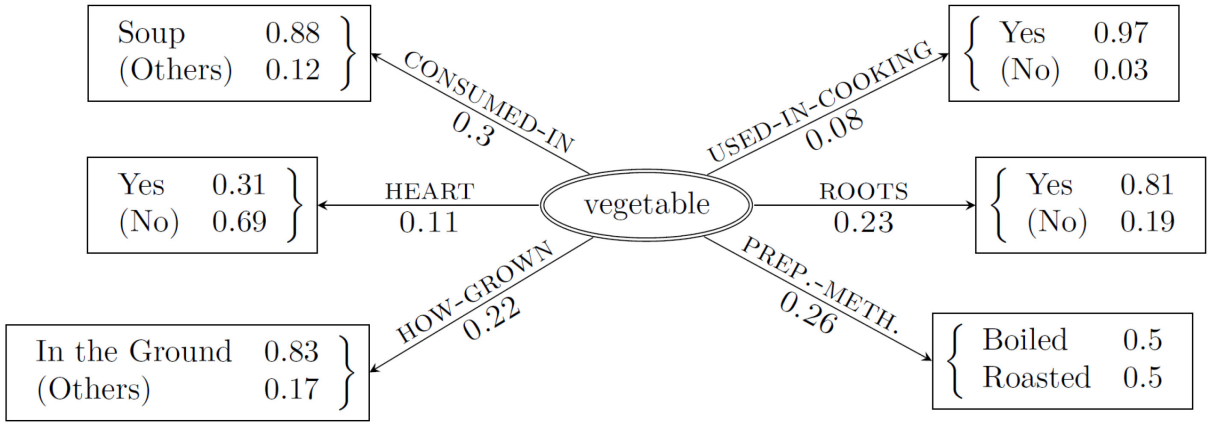

Figure 61: Prototype frame of vegetables resulting from experiment data.

Using only category properties and normalising diagnosticity as two scales from 0 to 5 corresponds to the view that being very probably a fruit based on the property in question means being not at all diagnostic for vegetables. To see how including all properties (also the contrast category properties) changes the predictions, diagnosticity was also calculated considering the whole 11-point-scale as relevant for both categories and normalised as:

$$\mathrm{Pr}_{\mathrm{norm_{all}}}(\mathrm{fruit}|\mathrm{V_{ij}}) = \frac{\mathrm{Pr}(\mathrm{fruit}|\mathrm{V_{ij}}) + 5}{10},$$

$$\mathrm{Pr}_{\mathrm{norm_{all}}}(\mathrm{vegetable}|\mathrm{V_{ij}}) = \frac{\left((-1) \cdot \mathrm{Pr}(\mathrm{fruit}|\mathrm{V_{ij}})\right) + 5}{10}.$$

Then, properties which are highly diagnostic for vegetables receive a very low diagnosticity for fruit and vice versa. A mean rating of -4 would for example be normalised to 0.9 for vegetables and 0.1 for fruit, meaning that having a similarity to the prototype in this attribute contributes largely to the typicality of vegetables and only slightly to the typicality of fruit. Like in the version with category properties only, the maximum per attribute was then determined and divided by the sum product of the category value probabilities with the attribute weights. The results of these calculations are in Table 50 for *fruit* and Table 51 for *vegetable*.

Table 50: Value probability and diagnosticity calculations for fruit with all properties.

property	attribute	value	mean Pr(V|C) rating	norm per A	mean Pr(C|P) rating	mean Pr(C|P) norm	max norm	diag A
will be eaten as dessert	CONSUMED-IN	Dessert	0.84	0.42	3.97	0.90	0.90	0.18
will be made into juice	CONSUMED-IN	Juice	0.93	0.47	2.07	0.71	0.90	0.18
will be eaten in soup	CONSUMED-IN	Soup	0.21	0.10	-3.67	0.13	0.90	0.18
has a heart	HEART	Yes	0.17	0.17	-1.33	0.37	0.37	0.07
grew on a tree	HOW-GROWN	Trees	0.86	0.44	2.23	0.72	0.72	0.15
grew on a plant	HOW-GROWN	On Plants	0.72	0.37	1.13	0.61	0.72	0.15
grew in the ground	HOW-GROWN	In the Ground	0.36	0.19	-2.73	0.23	0.72	0.15
is juicy	JUICINESS	Juicy	0.88	0.88	2.70	0.77	0.77	0.16
has sugar	MAIN-NUTRITIONAL-COMPONENT	Sugar	0.92	0.92	2.37	0.74	0.74	0.15
will be boiled	PREPARATION-METHOD	Boiled	0.31	0.31	-2.63	0.24	0.24	0.05
will be roasted	PREPARATION-METHOD	Roasted	0.25	0.25	-3.17	0.18	0.24	0.05
has roots	ROOTS	Yes	0.35	0.35	-2.80	0.22	0.22	0.04
has pips/seeds	SEEDS	Yes	0.88	0.88	2.13	0.71	0.71	0.14
is sweet	TASTE	Sweet	0.88	0.60	2.70	0.77	0.77	0.16
tastes tart	TASTE	Tart	0.59	0.40	2.50	0.75	0.77	0.16
will be used in cooking	USED-IN-COOKING	Yes	0.72	0.72	-1.03	0.40	0.40	0.08

Table 51: Value probability and diagnosticity calculations for vegetables with all properties.

property	attribute	value	mean Pr(V|C) rating	norm per A	mean Pr(C|P) rating	mean Pr(C|P) norm	max norm	diag A
will be eaten in soup	CONSUMED-IN	Soup	0.88	0.55	-3.67	0.87	0.87	0.20
will be made into juice	CONSUMED-IN	Juice	0.55	0.34	2.07	0.29	0.87	0.20
will be eaten as dessert	CONSUMED-IN	Dessert	0.17	0.10	3.97	0.10	0.87	0.20
has a heart	HEART	Yes	0.31	0.31	-1.33	0.63	0.63	0.15
grew in the ground	HOW-GROWN	In the Ground	0.83	0.45	-2.73	0.77	0.77	0.18
grew on a plant	HOW-GROWN	On Plants	0.69	0.37	1.13	0.39	0.77	0.18
grew on a tree	HOW-GROWN	Trees	0.32	0.17	2.23	0.28	0.77	0.18
is juicy	JUICINESS	Juicy	0.34	0.34	2.70	0.23	0.23	0.05
has sugar	MAIN-NUTRITIONAL-COM	Sugar	0.42	0.42	2.37	0.26	0.26	0.06
will be roasted	PREPARATION-METHOD	Roasted	0.83	0.50	-3.17	0.82	0.82	0.19
will be boiled	PREPARATION-METHOD	Boiled	0.81	0.50	-2.63	0.76	0.82	0.19
has roots	ROOTS	Yes	0.81	0.81	-2.80	0.78	0.78	0.18
has pips/seeds	SEEDS	Yes	0.45	0.45	2.13	0.29	0.29	0.07
tastes tart	TASTE	Tart	0.28	0.28	2.50	0.25	0.25	0.06
is sweet	TASTE	Sweet	0.37	0.37	2.70	0.23	0.25	0.06
will be used in cooking	USED-IN-COOKING	Yes	0.97	0.97	-1.03	0.60	0.60	0.14

The frames with all properties are in Figure 62 for fruit and Figure 63 for vegetables. Now, both categories are represented with two attributes with three values, two attributes with two values and six attributes with one value.

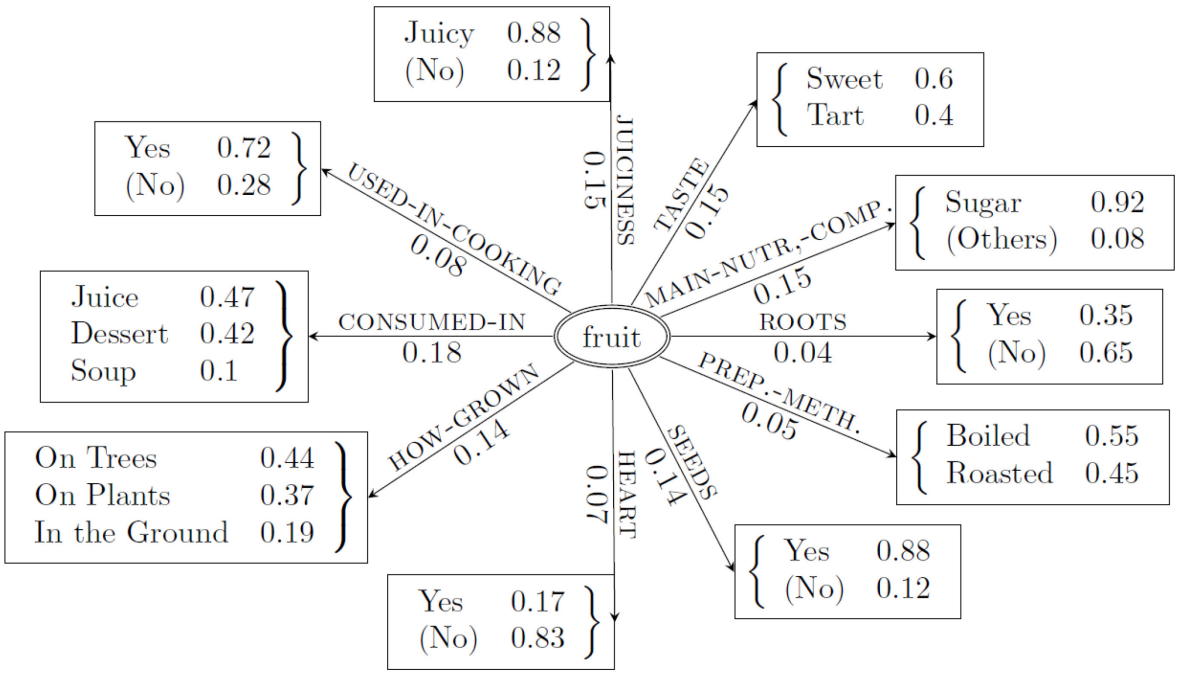

Figure 62: Frame representation of fruit with all properties with attribute complementing values in brackets.

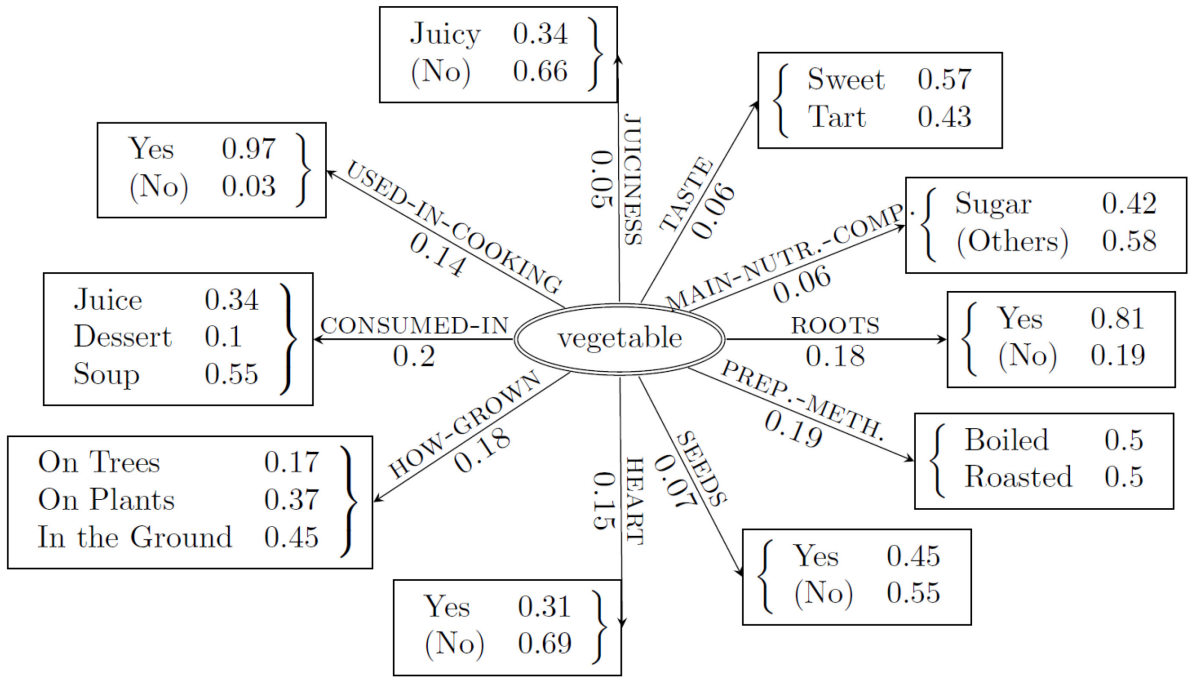

Figure 63: Frame representation of vegetables with all properties with attribute complementing values in brackets.

The SC frames only include value probability information and were normalised like the C frames: if the sum of the mean probabilities Pr ($V_{ij}|C$) for an attribute was greater than 1, all probabilities were divided by the sum of the corresponding attribute. Value probabilities with sums smaller than 1 per attribute were left as they were.

7.1.2 **Predictions and results**

Typicality predictions were calculated with the above-mentioned formula and the input described in section 7.1.1. The similarities, that is the sum of the minimal value probabilities of SC and C for all attributes, were multiplied with the attribute diagnosticities. The sum of all diagnosticity-weighted similarities is the predicted typicality. Figure 64 shows an example for the calculation of *apple*'s typicality for *fruit* with 3 attributes.

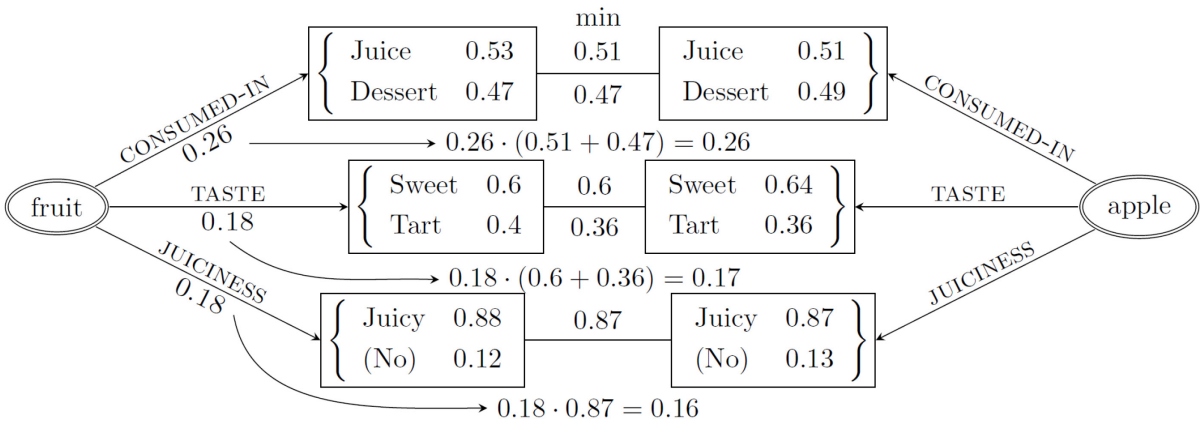

Figure 64: Calculation of the typicality contribution of 3 attributes for apple.

Due to the high intersubjective variability in the probability judgments for one third of the SC-property-pairs for *fruit* and almost half of the pairs for *vegetable* (detailed in section 6.3.5.2 and discussed in section 6.3.5.3), the predictions were additionally calculated in a cleaned version in which only those SC-property-pairs with an IQR smaller or equal to 40 were used. For cleaned calculations, it was necessary to normalise diagnosticities with the sum-product of the subset of properties that were included after cleaning, because the maximally possible similarity was different for each SC depending on which properties were excluded. Cleaning thus leads to the creation of a specific (sub-) prototype frame for each SC.

To assess the stability of the results, the jackknife delete-one-technique (see section 3.1.4) was used, from which the average standard error of the correlation can be estimated by averaging over the correlations resulting from successively leaving one datapoint out.

The Pearson correlations between mean rated and predicted typicality as well as the SEs and 95% confidence intervals from jackknifing are in Table 52 for *fruit* and Table 53 for *vegetable*. For both categories, the highest predictions are found for cleaned data with category properties only: they are .84 ($p<.001$, CI: [0.75, 0.94]) for *fruit* and .68 ($p<.001$, CI: [0.4, 0.96]) for *vegetable*. All predictions show a clear relationship with the ratings with high effect sizes ≥0.5 and low p-values ≤0.01, except for the predictions with all properties and uncleaned data for *vegetable* (.46, $p<0.05$, CI: [0.06. 0.86]). The cleaned versions have higher effect sizes (+.17 for *fruit* and +15 for *vegetable*), lower p-values and smaller confidence intervals. The predictions for *fruit* have higher correlations than those for *vegetable* (+.14 for all data, +16 for cleaned data).[33]

[33] The exclusion of the four control properties changed the correlations as follows: for *fruit* with category properties and all data: +.08, cleaned data -.01, with all properties and all data: +.14, cleaned data +.01. For *vegetable* with category properties and all data: -.02, cleaned data: -.01, with all properties and all data: -.02, cleaned data: -.03.

Table 52: Pearson correlations between predicted and observed typicality and jackknife results for fruit.

properties	datatype	*r*	p	SE	CI
category	all	.67	.001	.144	[0.39, 0.96]
	cleaned	.84	.000	.050	[0.75, 0.94]
all	all	.60	.003	.182	[0.25, 0.96]
	cleaned	.78	.000	.060	[0.66, 0.89]

Table 53: Pearson correlations between predicted and observed typicality and jackknife results for vegetables.

properties	datatype	*r*	p	SE	CI
category	all	.51	.015	.165	[0.19, 0.83]
	cleaned	.67	.001	.145	[0.38, 0.95]
all	all	.45	.037	.205	[0.05, 0.85]
	cleaned	.59	.004	.162	[0.27, 0.9]

Plots of the linear regressions corresponding to these results are in Figure 65 for *fruit*. They show that the models with all data cover the high typicality range, but the regression line is not steep enough, i.e., the models predict typicality for SCs with low typicality too high. With cleaned data, the SCs in the lower typicality intervals are predicted better, in particular with category properties only.

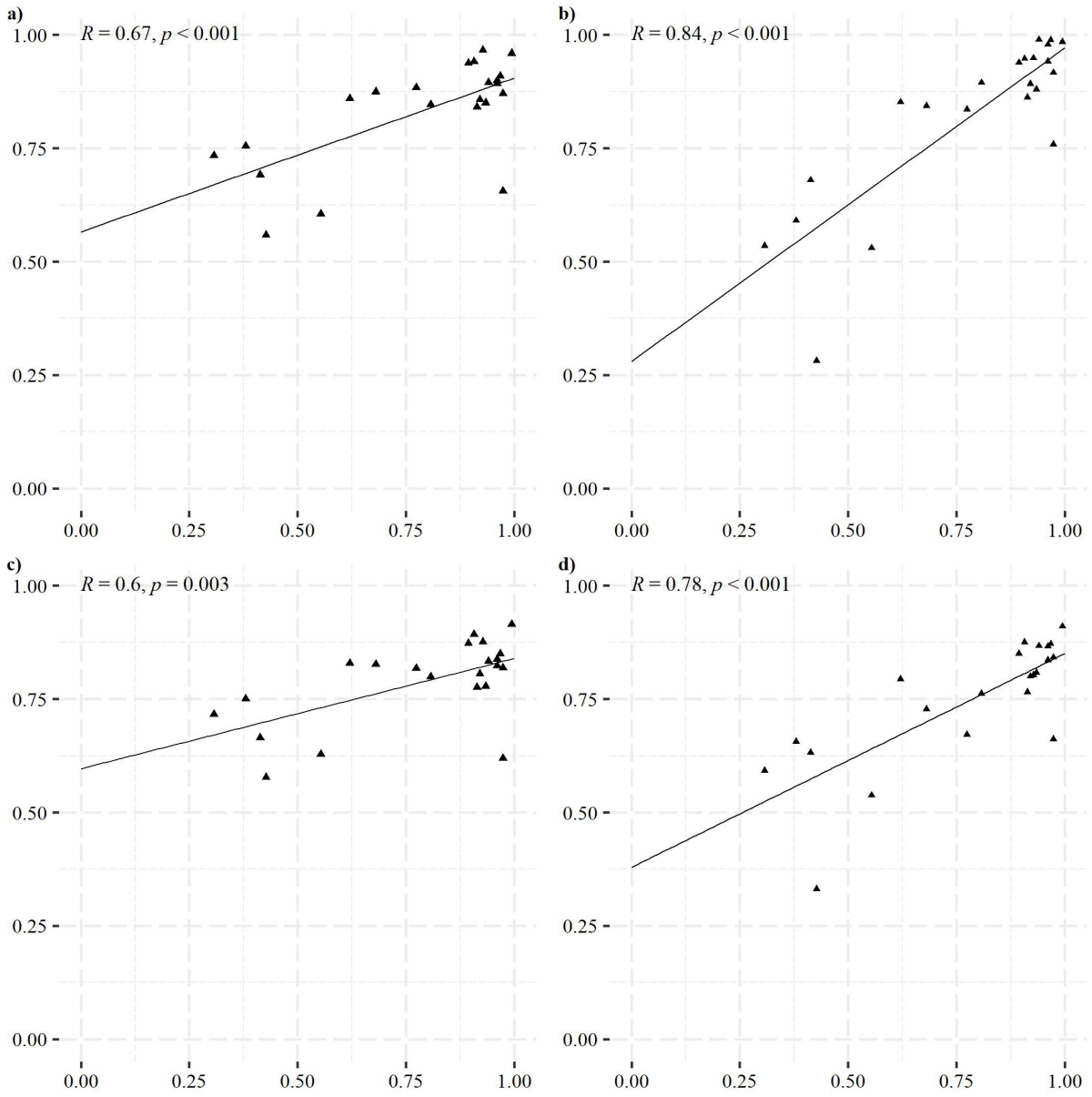

Figure 65: Linear regression plots for fruit typicality predictions with regression line and predicted typicality (triangles) with category properties and a) all data, b) cleaned data, and with all properties and c) all data, d) cleaned data.

The regression plots for *vegetable* are in Figure 66. The predictions do not capture the typicality ordering well, except for the version with cleaned data and category properties only, where the tendency of the predictions is right for most SCs, except for several SCs with a rated typicality between .73 and .77, whose typicality is predicted too low or too high. With contrast properties and uncleaned data, the regression line is almost flat. In the cleaned version with contrast properties, the predictions are too low to capture highly typical SCs.

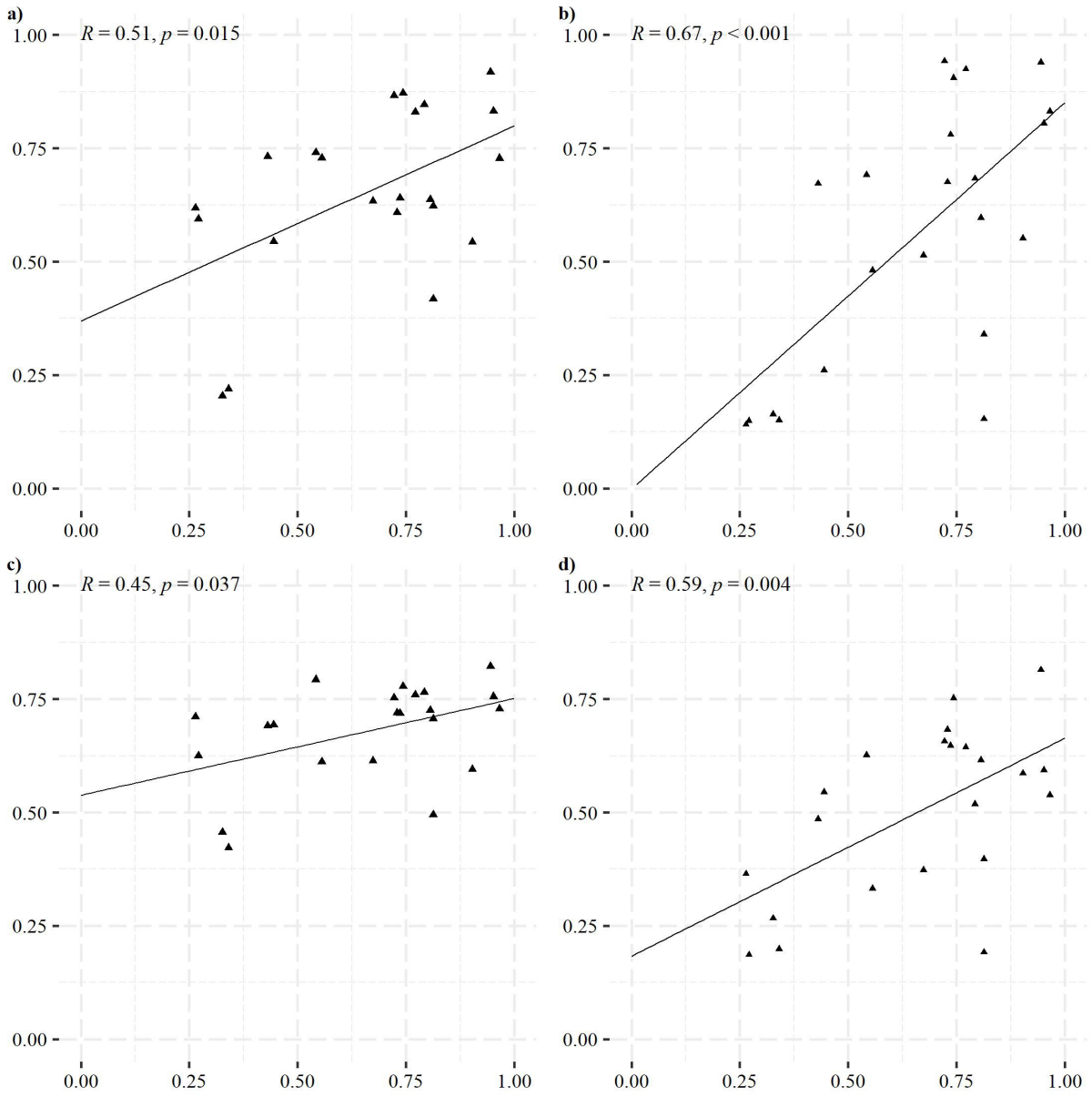

Figure 66: Linear regression plots for vegetable typicality predictions with regression line and predicted typicality (triangles) with category properties (top) and all properties (bottom) for a), c) with all data, b), d) with cleaned data.

The rated typicality compared to the predicted typicality for all SCs and for the predictions without contrast category properties is in Table 54, including the difference it makes to leave each SC out, computed in the jackknife procedure. For *fruit*, changes greater than |0.05| are observed for *banana* and *avocado* for the calculations with all data: leaving *banana* out raises the correlation substantially (+.13) and leaving *avocado* out lowers it (-.06). The predictions with cleaned data are very stable with no difference greater than |0.3|. For *vegetable*, leaving out *lettuce* raises the correlation (+.06). Three SCs stabilise the correlation and reduce it when left out when all data are used: *carrot* (-.06), *avocado* and *pickle* (-.1). For cleaned data, leaving *lettuce* out again improves the correlation (+.08).

For *fruit*, the range (maximum minus minimum) of predictions with all data is .41, while the range of typicality is .69. The predictions with cleaned data have a more similar range of .71. For *vegetable*, the predictions cover the same range (.71) as the typicality data (.69). In the cleaned version they have a slightly higher range (.80) than the one observed for typicality.

Table 54: Normalised mean rated and predicted typicality with all data and cleaned data and changes in correlation when leaving each SC out for a) fruit and b) vegetables.

a)

subcategory	mean rated	mean predicted all	mean predicted cleaned	jackknife all	jackknife cleaned
apple	.99	.96	.99	-.02	-.01
banana	.97	.66	.76	.13	.03
strawberry	.97	.87	.92	0	0
peach	.97	.91	.99	-.01	-.01
grape	.96	.90	.94	-.01	0
pineapple	.96	.89	.98	-.01	-.01
mango	.94	.89	.99	-.01	0
blueberry	.93	.85	.88	0	0
passion fruit	.93	.97	.95	-.01	0
blackberry	.92	.86	.89	0	0
watermelon	.91	.84	.86	0	0
plum	.91	.94	.95	-.01	0
pomegranate	.89	.94	.94	0	0
lime	.81	.85	.89	0	0
papaya	.77	.88	.84	0	0
prune	.68	.87	.84	.01	.01
fig	.62	.86	.85	.02	.02
rhubarb	.55	.61	.53	-.02	-.01
avocado	.43	.56	.28	-.06	.02
coconut	.41	.69	.68	-.03	.01
tomato	.38	.76	.59	0	-.01
pumpkin	.31	.73	.53	0	-.02

b)

subcategory	mean rated	mean predicted all	mean predicted cleaned	jackknife all	jackknife cleaned
broccoli	.97	.73	.83	-.01	-.03
cauliflower	.95	.83	.81	-.04	-.03
carrot	.94	.92	.94	-.06	-.04
pea	.90	.54	.55	.04	.01
spinach	.81	.62	.34	0	.03
lettuce	.81	.42	.15	.06	.08
zucchini	.81	.64	.60	-.01	-.01
onion	.79	.85	.68	-.03	-.02
green onion	.77	.83	.92	-.02	-.01
sweet potato	.74	.87	.91	-.02	-.01
corn	.74	.64	.78	-.01	-.01
eggplant	.73	.61	.68	-.01	-.02
potato	.72	.87	.94	-.01	0
radish	.67	.63	.51	-.01	-.01
mushroom	.56	.73	.48	0	-.02
pumpkin	.54	.74	.69	0	.01
tomato	.44	.54	.26	-.03	-.03
garlic	.43	.73	.67	.02	.03
pickle	.34	.22	.15	-.10	-.05
avocado	.33	.20	.16	-.10	-.05
parsley	.27	.59	.15	0	-.06
rhubarb	.26	.62	.14	.01	-.06

Figure 67 and Figure 68 show the typicality contributions of each property with

$$\text{typcon}(\text{P}|\text{SC}) = \text{diag}(\text{A}_\text{i}) \cdot \min(\Pr(\text{P}|\text{SC}), \Pr(\text{P}|\text{C}))$$

in stacked barplots for raw and cleaned data for *fruit* resp. *vegetable*. It can be seen well that cleaning, which means focusing on those properties about whose probability the subjects showed more agreement, improves the typicality prediction. For *fruit*, the highly typical SCs are predicted very accurately in all versions, except for *banana*, which is consistently predicted too low. On the low typicality level, predictions tend to be too high, except for *rhubarb*, whose typicality is well predicted in all versions. The predictions for *avocado*, *pumpkin* and *tomato* are lowered to a more corresponding level in the versions with cleaned data, while *coconut* is predicted too high in all models.

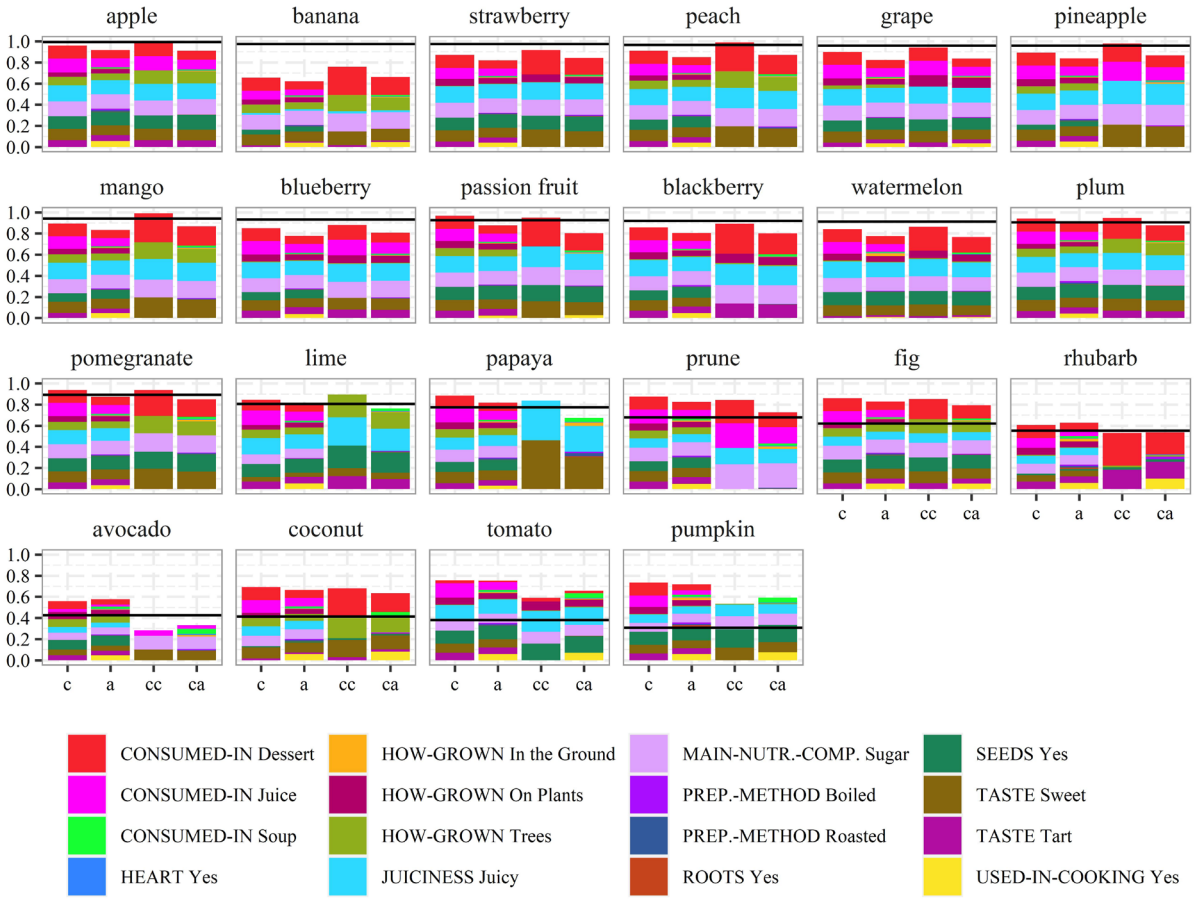

Figure 67: Typicality contribution of each property per attribute-value-pair with category properties ("c") and all properties ("a") for all data and cleaned data ("cc" resp. "ca") with rated typicality (horizontal line) for fruit.

For *vegetable*, the highly typical SCs are predicted too low, except for *carrot*, whose predictions are accurate in both the cleaned and uncleaned version with category properties only. Many predictions for SCs on the medium typicality level are too high. On the low typicality level, the predictions are often too high with uncleaned data and too low with cleaned data.

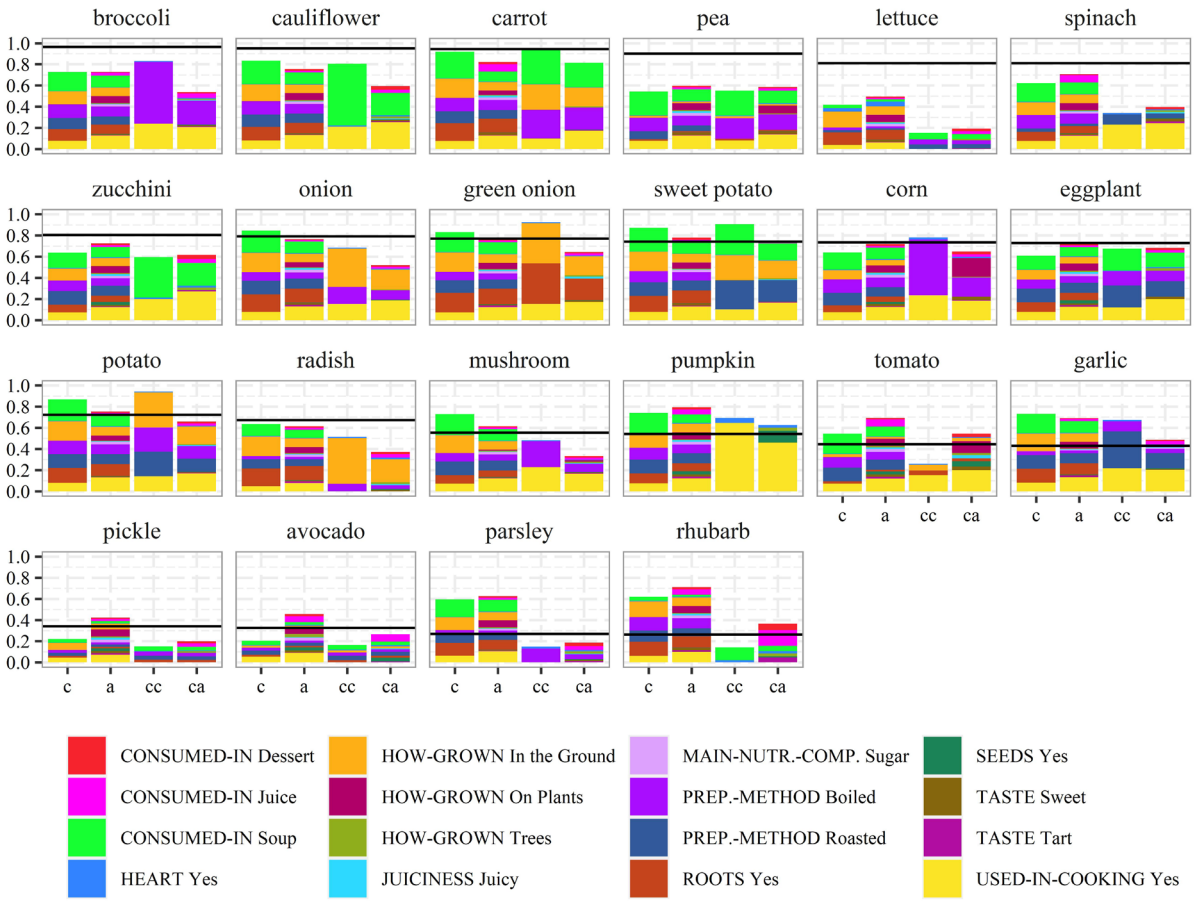

Figure 68: Typicality contribution of each property per attribute-value-pair with category properties ("c") and all properties ("a") for all data and cleaned data ("ca" resp. "cc") with rated typicality (horizontal line) for vegetables.

What this analysis shows is that the constructed models capture well what makes very typical fruit very typical and what makes some medium typical vegetables medium typical, but the other typicality levels are covered less well.

7.1.3 **Alternative models**

In the probabilistic prototype model, diagnosticity weights are applied on the attribute level to account for the fact that having one very diagnostic value in an attribute should boost the typicality contribution for all values of that attribute. Two alternative ways to predict typicality were computed in which diagnosticity was either applied on the property level or not at all.

To apply diagnosticity weights on the property level, the maximum per attribute is not computed and the diagnosticities are directly $\Pr(C|V_{ij}) = \Pr(C|P_k)$, divided by the dot product of the category property probabilities with the attribute weights to normalise the predictions between 0 and 1:

$$\mathrm{diag}(P_k|C) = \frac{\Pr(C|P_k)}{\sum_{k=1}^{l} \Pr(C|P_k)} \cdot \frac{1}{\sum_{k=1}^{l} \Pr(P_k|C)}.$$

The calculated values compared to the attribute diagnosticities are in Table 55 for category properties and Table 56 for the version with contrast properties. The absolute values tend to be a bit lower because the value frequencies have higher values when they are not normalised per

attribute and thus contribute more to the equation. For example, the mean value probability rating for *will be made into juice* is normalised .93 and normalised per attribute .53 for *fruit*.

Table 55: Diagnosticity applied to the attribute level and property level for a) fruit and b) vegetables for category properties.

a)

property	mean Pr(V|C) rating	norm per A	mean Pr(C|P) rating	mean Pr(C|P) norm	diag A	diag P
will be made into juice	.93	.53	2.1	.41	.26	.11
will be eaten as dessert	.84	.47	4.0	.79	.26	.22
grew on a tree	.86	.54	2.2	.45	.15	.12
grew on a plant	.72	.46	1.1	.23	.15	.06
is juicy	.88	.88	2.7	.54	.18	.15
has sugar	.92	.92	2.4	.47	.15	.13
has pips/seeds	.88	.88	2.1	.43	.14	.12
is sweet	.88	.60	2.7	.54	.18	.15
tastes tart	.59	.40	2.5	.50	.18	.14

b)

property	mean Pr(V|C) rating	norm per A	mean Pr(C|P) rating	mean Pr(C|P) norm	diag A	diag P
will be eaten in soup	.88	.88	-3.7	.73	.30	.26
has a heart	.31	.31	-1.3	.27	.11	.10
grew in the ground	.83	.83	-2.7	.55	.22	.20
will be roasted	.83	.50	-3.2	.63	.26	.23
will be boiled	.81	.50	-2.6	.53	.26	.19
has roots	.81	.81	-2.8	.56	.23	.20
will be used in cooking	.97	.97	-1.0	.21	.08	.07

Table 56: Diagnosticity applied to the attribute level and property level for a) fruit and b) vegetables for all properties.

a)

property	mean Pr(V|C) rating	norm per A	mean Pr(C|P) rating	mean Pr(C|P) norm	diag A	diag P
will be eaten as dessert	.84	.42	4.0	.90	.18	.14
will be made into juice	.93	.47	2.1	.71	.18	.11
will be eaten in soup	.21	.10	-3.7	.13	.18	.02
has a heart	.17	.17	-1.3	.37	.07	.06
grew on a tree	.86	.44	2.2	.72	.15	.12
grew on a plant	.72	.37	1.1	.61	.15	.10
grew in the ground	.36	.19	-2.7	.23	.15	.04
is juicy	.88	.88	2.7	.77	.16	.12
has sugar	.92	.92	2.4	.74	.15	.12
will be boiled	.31	.31	-2.6	.24	.05	.04
will be roasted	.25	.25	-3.2	.18	.05	.03
has roots	.35	.35	-2.8	.22	.04	.04
has pips/seeds	.88	.88	2.1	.71	.14	.11
is sweet	.88	.60	2.7	.77	.16	.12
tastes tart	.59	.40	2.5	.75	.16	.12
will be used in cooking	.72	.72	-1.0	.40	.08	.06

b)

property	mean Pr(V|C) rating	norm per A	mean Pr(C|P) rating	mean Pr(C|P) norm	diag A	diag P
will be eaten in soup	.88	.55	-3.7	.87	.20	.17
will be made into juice	.55	.34	2.1	.29	.20	.06
will be eaten as dessert	.17	.10	4.0	.10	.20	.02
has a heart	.31	.31	-1.3	.63	.15	.12
grew in the ground	.83	.45	-2.7	.77	.18	.15
grew on a plant	.69	.37	1.1	.39	.18	.08
grew on a tree	.32	.17	2.2	.28	.18	.05
is juicy	.34	.34	2.7	.23	.05	.04
has sugar	.42	.42	2.4	.26	.06	.05
will be roasted	.83	.50	-3.2	.82	.19	.16
will be boiled	.81	.50	-2.6	.76	.19	.15
has roots	.81	.81	-2.8	.78	.18	.15
has pips/seeds	.45	.45	2.1	.29	.07	.06
tastes tart	.28	.28	2.5	.25	.06	.05
is sweet	.37	.37	2.7	.23	.06	.04
will be used in cooking	.97	.97	-1.0	.60	.14	.12

In order to represent property diagnosticities within the frame model, the properties are regarded as binary attributes (their probabilities relating to their value 1 or “true”). The resulting frames

are in Figure 69 for *fruit* and Figure 70 for *vegetable* for the models with category properties only and in Figure 71 for *fruit* and Figure 72 for *vegetable* with contrast properties.

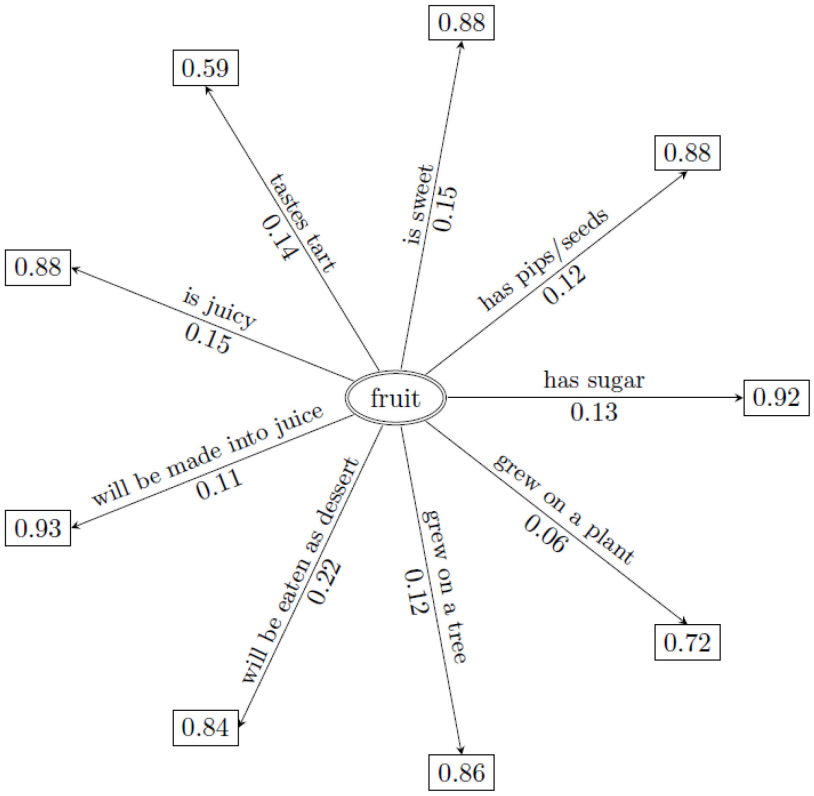

Figure 69: Frame representation of fruit with category properties and diagnosticity on the property level.

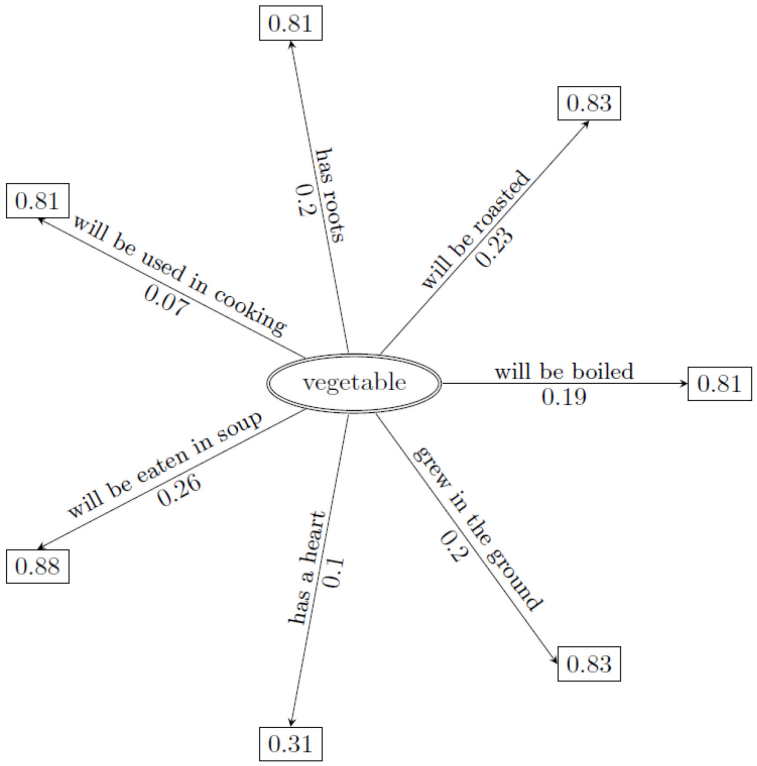

Figure 70 Frame representation of vegetables for category properties and diagnosticity on the property level.

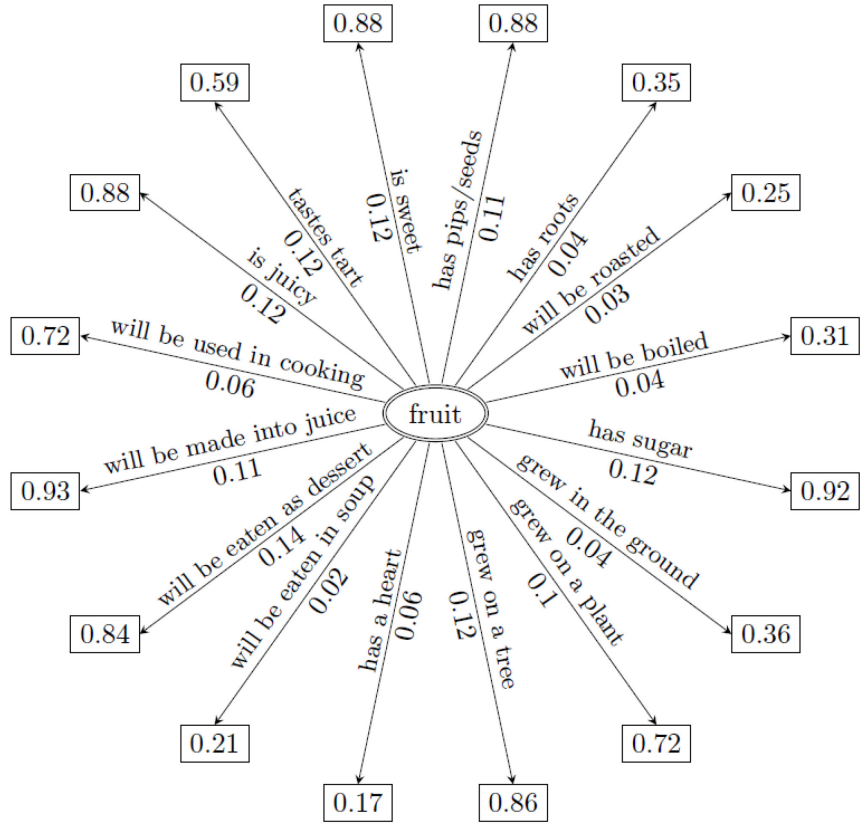

Figure 71: Frame representation for fruit with all properties and diagnosticity on the property level.

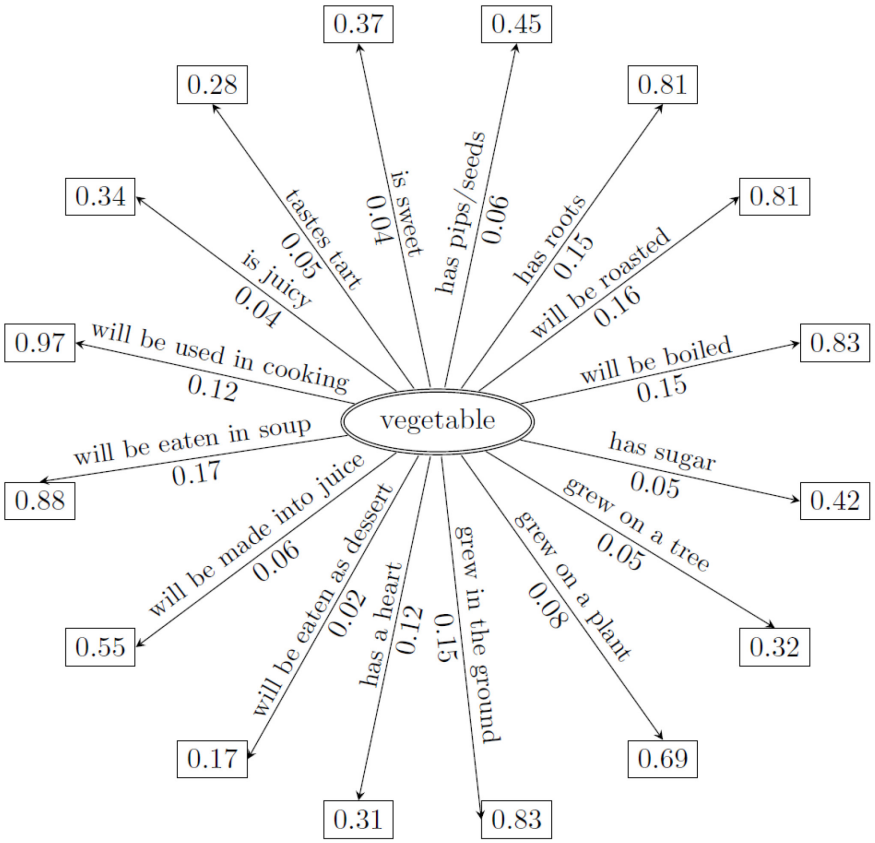

Figure 72: Frame representation of vegetables with all properties and diagnosticity on the property level.

The correlations of the predictions with typicality for *fruit* are in Table 57 and the regression plots in Figure 73. The predictions with cleaned data are in the same range as when diagnosticity is applied to the attribute level, with slightly lower SEs (-.1 for category properties and -.15 for contrast properties). Without cleaning, they are much higher with +.18 for category properties and +.19 for contrast properties. The high difference between the different sets of properties and between cleaned and uncleaned data almost disappear in this constellation. This is also visible from the regression plots. Here, a slight qualitative superiority of the cleaned versions is visible: the predictions with cleaned data have a higher range (steeper regression line) and capture the absolute values of SCs with low typicality better.

Table 57: Pearson correlations between predicted and observed typicality and jackknife results for fruit for diagnosticity on the property level.

properties	datatype	*r*	p	SE	CI
category	all	.85	.000	.064	[0.73, 0.97]
	cleaned	.87	.000	.040	[0.79, 0.95]
all	all	.79	.000	.087	[0.62, 0.96]
	cleaned	.81	.000	.045	[0.72, 0.9]

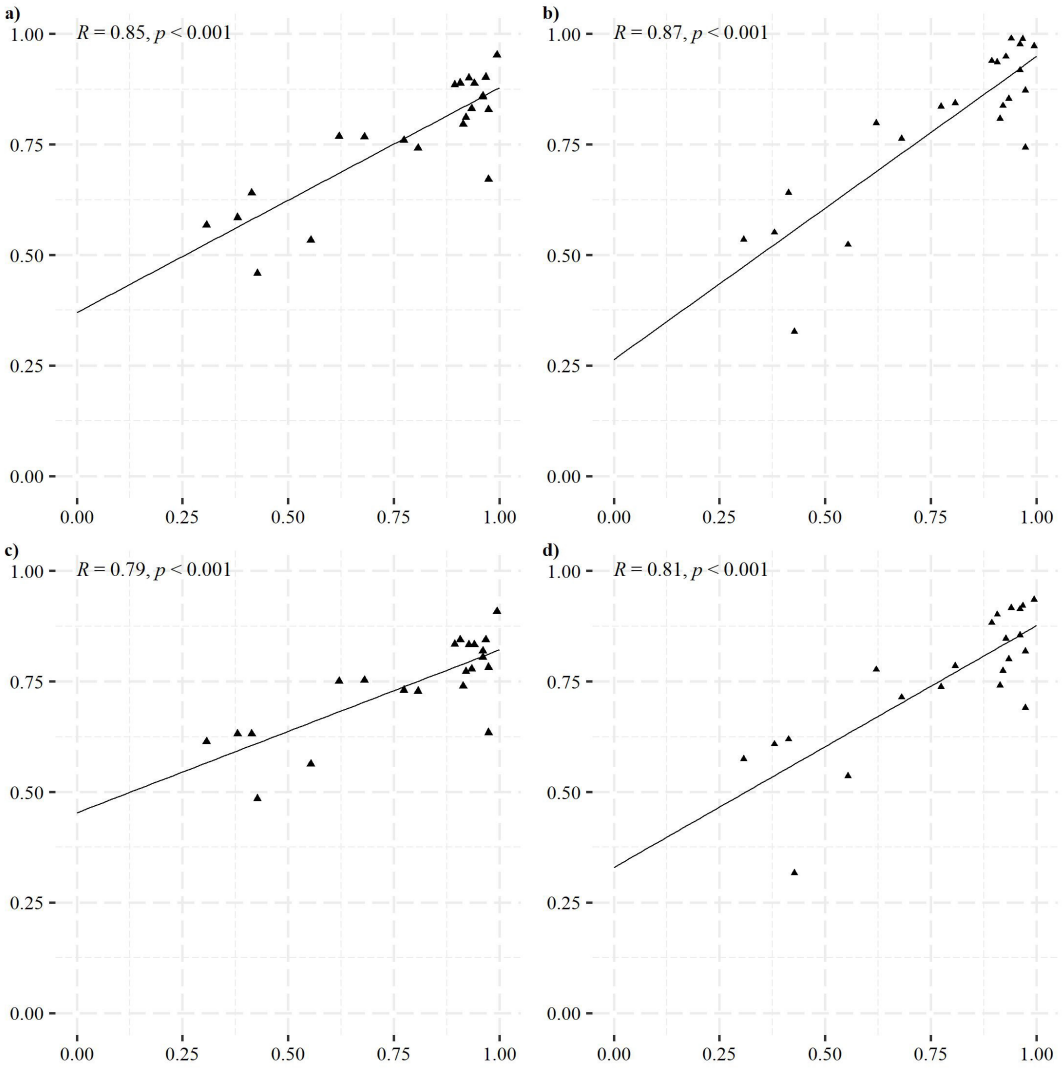

Figure 73: Linear regression plots for fruit typicality predictions with diagnosticity weights on the property level with regression line and predicted typicality (triangles) with category properties and a) with all data, b) with cleaned data, and with all properties and c) all data, d) cleaned data.

The fact that the predictions with uncleaned data are much higher was investigated closer. Table 58 shows the differences in prediction between the attribute and property models with category properties. They differ most for SCs with a mean rated typicality <.89, where the property model predicts the lower typicality of *tomato* and *pumpkin* (-.17), *papaya* (-.12), *prune* (-.11) and *avocado* and *lime* (-.10) more accurately. Figure 74 shows the typicality contributions for each attribute-value-pair for these SCs. It can be seen that the lower, and therefore more accurate, predictions are due to giving a smaller weight to *will be made into juice*, *is juicy* and *grew on plants*, which have a higher diagnosticity in the attribute model. Similar results were found for the models with contrast properties.

Table 58: Rated and predicted mean for attribute and property models with category properties and all data and the difference in predictions for fruit.

subcategory	mean rated	mean predicted attribute	mean predicted property	diff
apple	.99	.96	.95	.01
banana	.97	.66	.67	.02
strawberry	.97	.87	.83	.04
peach	.97	.91	.90	.01
grape	.96	.90	.86	.04
pineapple	.96	.89	.86	.03
mango	.94	.89	.89	.01
blueberry	.93	.85	.83	.02
passion fruit	.93	.97	.90	.07
blackberry	.92	.86	.81	.05
watermelon	.91	.84	.80	.04
plum	.91	.94	.89	.05
pomegranate	.89	.94	.89	.05
lime	.81	.85	.74	.10
papaya	.77	.88	.76	.12
prune	.68	.87	.77	.11
fig	.62	.86	.77	.09
rhubarb	.55	.61	.53	.07
avocado	.43	.56	.46	.10
coconut	.41	.69	.64	.05
tomato	.38	.76	.58	.17
pumpkin	.31	.73	.57	.17

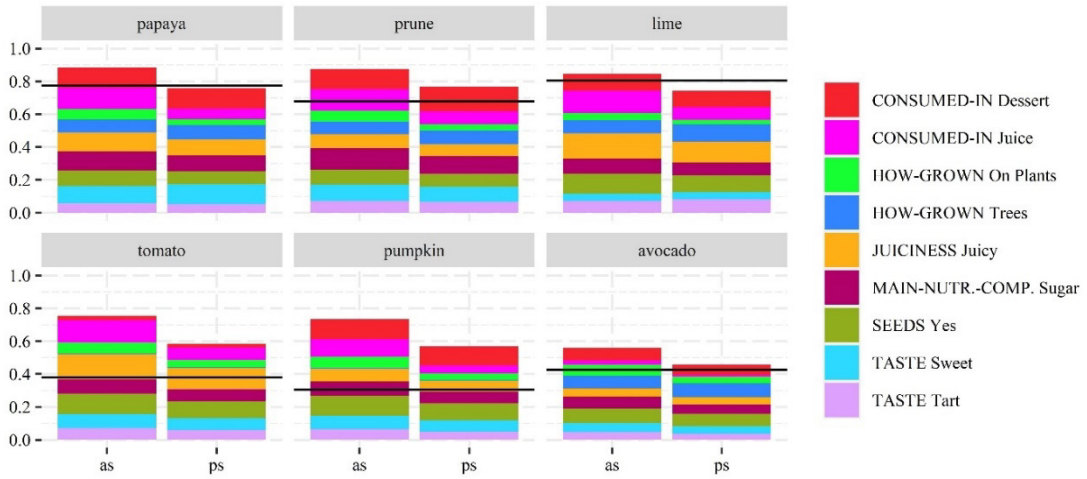

Figure 74: Typicality contributions for each attribute value for the predictions of the attribute model ("as") and the property model ("ps") for category properties with all data for the 6 fruit SCs with the highest difference in prediction.

In Table 59 are the correlations and in Figure 75 the regression plots of the predictions of the property model for *vegetable*. The correlations are in the same range as the attribute versions. The version with contrast properties and uncleaned data is higher by +.06 and has a lower p-value. The smaller difference to the attribute model is due to the fact that the frame for *vegetable* with category properties contains only one attribute with more than one value which leads to almost identical predictions for both models.

Table 59: Pearson correlations between predicted and observed typicality and jackknife results for vegetables.

properties	datatype	*r*	p	SE	CI
category	all	.53	.011	.157	[0.22, 0.84]
	cleaned	.68	.001	.144	[0.39, 0.96]
all	all	.51	.015	.168	[0.18, 0.84]
	cleaned	.62	.002	.166	[0.3, 0.95]

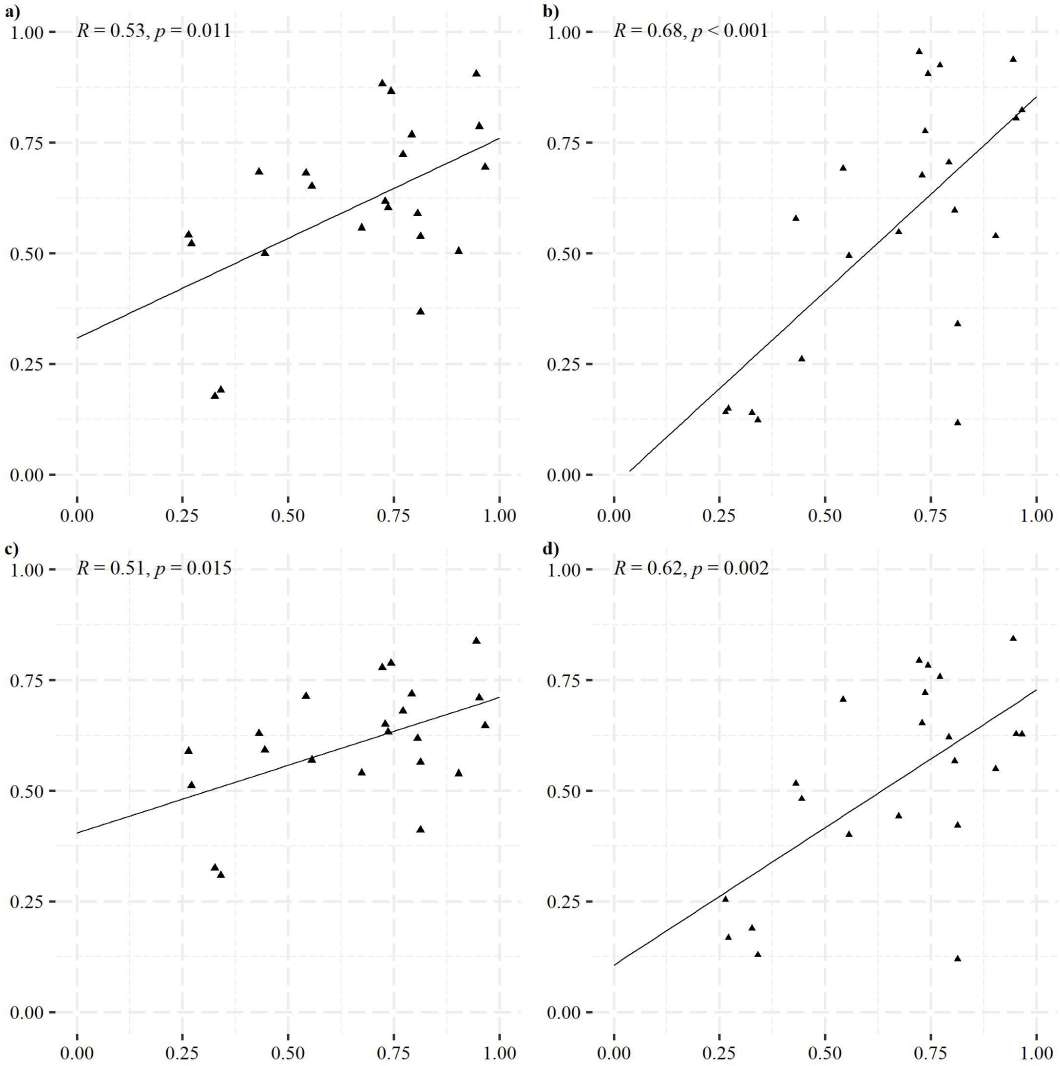

Figure 75: Linear regression plots for vegetable typicality predictions with diagnosticity weights on the property level with regression line and predicted typicality (triangles) with category properties and a) with all data, b) with cleaned data, and with all properties and c) all data, d) cleaned data.

We now turn to the second alternative model of computing typicalities. This model allows to judge the gain from including diagnosticity ratings in the predictions, because here they were completely omitted and only the similarities between the SC and C were summed up and divided by the sum of the maximally possible category probabilities either only for the l category properties:

$$\mathrm{typ}(SC, C) = \frac{\sum_{k=1}^{l} \mathrm{Sim}(SC, C|P_k)}{\sum_{k=1}^{l} \Pr(P_k|C)} \text{ for } l = \text{all } \Pr(C|P_k) > 0,$$

or for all k properties:

$$\mathrm{typ}(SC, C) = \frac{\sum_{k=1}^{16} \mathrm{Sim}(SC, C|P_k)}{\sum_{k=1}^{16} \Pr(P_k|C)}.$$

The frames correspond to the ones for the P models without diagnosticities. For *fruit*, the correlations are in Table 60 and the results of the linear regressions are in Figure 76. For the model with category properties and all data, the correlations are higher (+.15) than the model with diagnosticity weights on the attribute level and only slightly lower (-.03) than the one with diagnosticity weights on the property level. The model with all properties and cleaned data has lower correlations than the other two (-.08 for the attribute and -.11 for the property model).

Table 60: Pearson correlations for models without diagnosticity between predicted and observed typicality and jackknife results for fruit.

properties	datatype	*r*	p	SE	CI
category	all	.82	.000	.081	[0.66, 0.98]
	cleaned	.83	.000	.042	[0.75, 0.91]
all	all	.63	.002	.141	[0.35, 0.9]
	cleaned	.70	.000	.069	[0.57, 0.84]

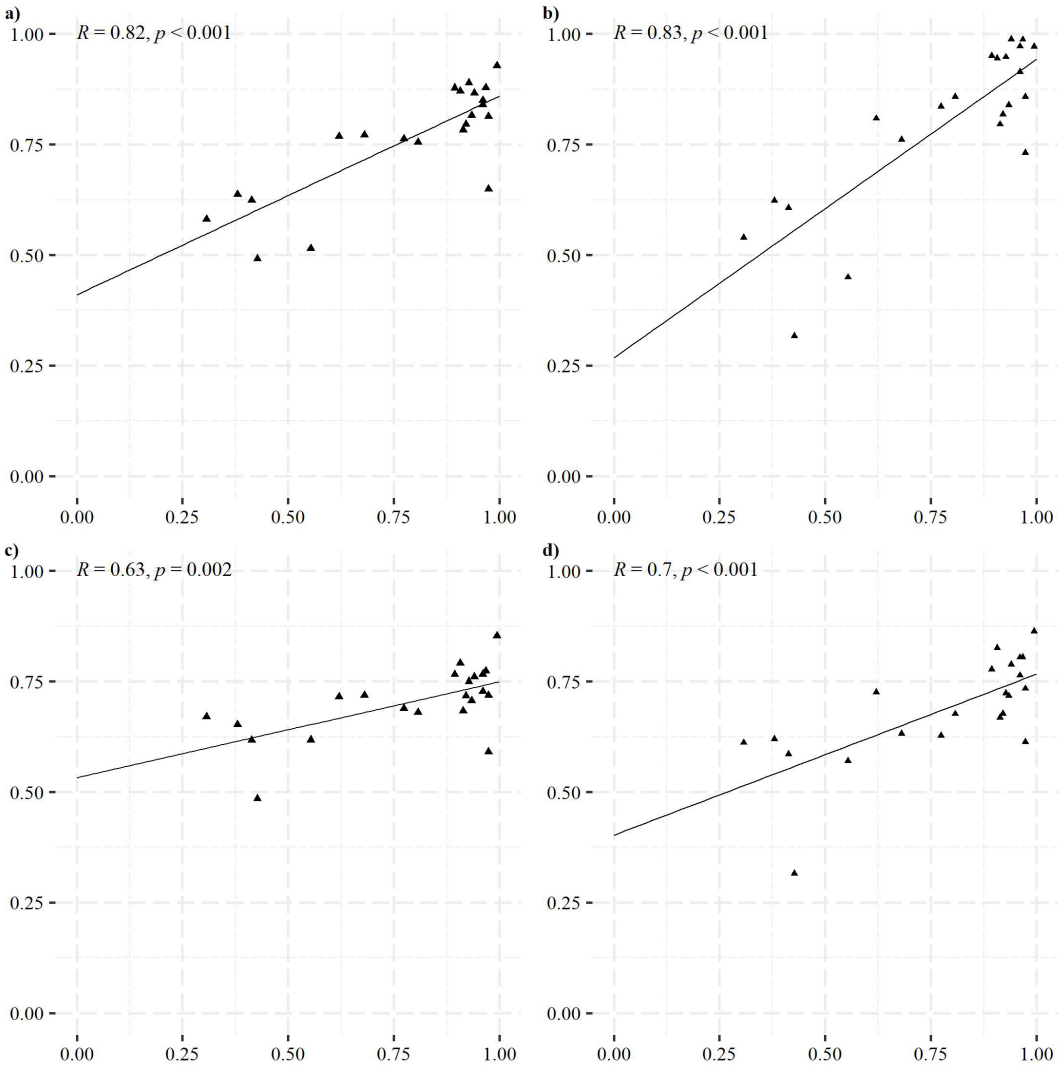

Figure 76: Linear regression plots for fruit typicality predictions without diagnosticity weights with regression line predicted typicality (triangles) with category properties and a) with all data, b) with cleaned data, and with all properties and c) all data, d) cleaned data.

For *vegetable*, the correlation results are in Table 61 and the regression plots in Figure 77. The models with category properties have correlations in the same range as the other models. With all properties, the correlations are lower, with all data -.07 for the attribute and -.13 for the property model and with cleaned data -.11 for the attribute and -.14 for the property model.

Table 61: Pearson correlations for models without diagnosticity between predicted and observed typicality and jackknife results for vegetables.

properties	datatype	*r*	p	SE	CI
category	all	.54	.009	.153	[0.24, 0.84]
	cleaned	.69	.000	.154	[0.38, 0.99]
all	all	.38	.085	.209	[-0.03, 0.78]
	cleaned	.48	.023	.196	[0.1, 0.87]

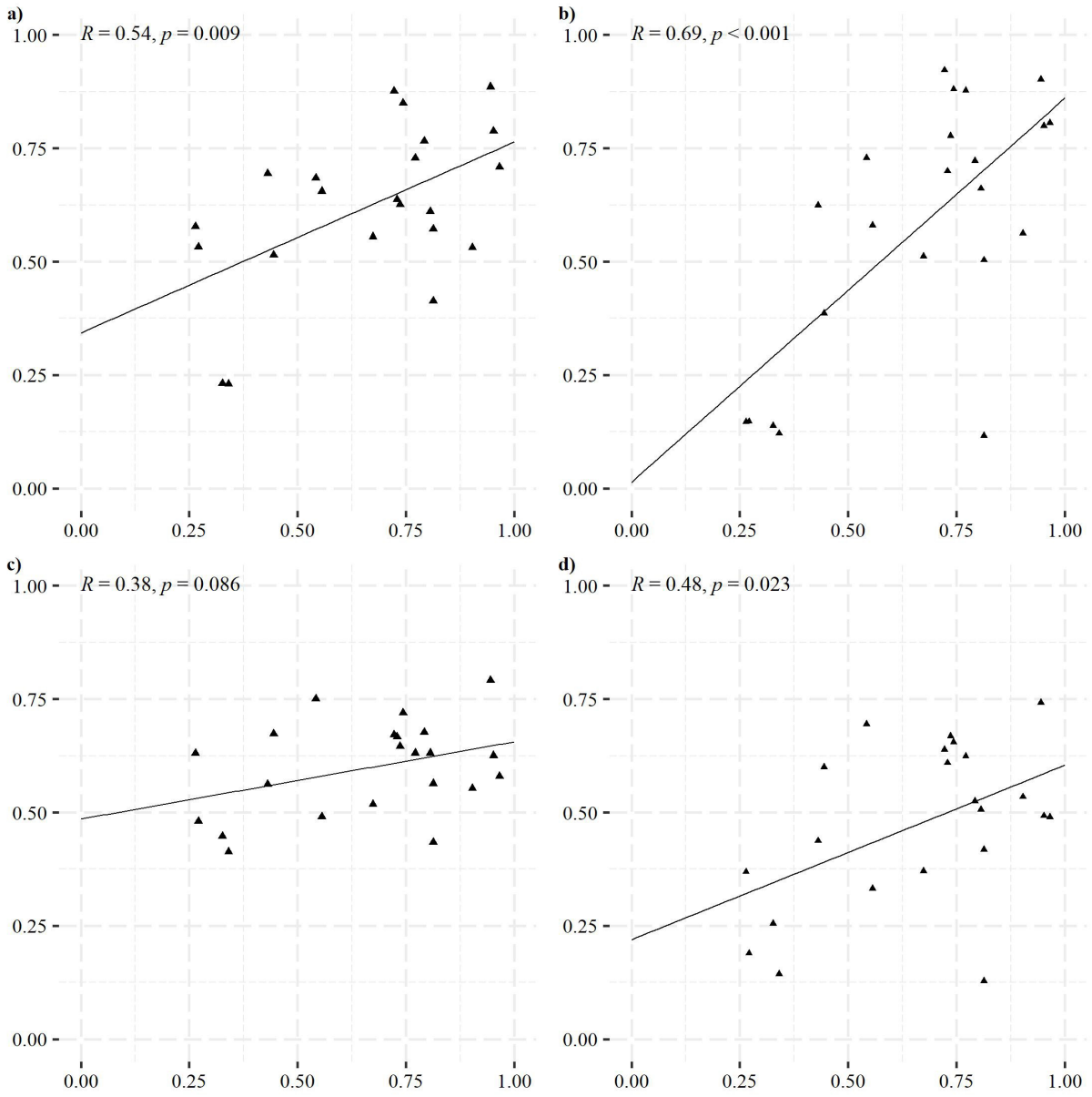

Figure 77: Linear regression plots for vegetable typicality predictions without diagnosticity weights with regression line and predicted typicality (triangles) with category properties and a) with all data, b) with cleaned data, and with all properties and c) all data, d) cleaned data.

A summary of the correlations between the model predictions and mean typicality are in Table 62 for both categories. For *fruit*, the correlations are high (between .6 and .87), and the p-values are low (all <0.01). For *vegetable*, the correlations are high except for the models with no diagnosticity weights and all properties and with attribute diagnosticity weights, all properties and uncleaned data. Correlations are between .39 and .70 and the p-values are ≤.01 except for two cases. For *fruit*, applying diagnosticity on the property level makes a mean difference of +.11 compared to the attribute predictions. For *vegetable*, the difference is low (+.03), which is to be expected as only one attribute has more than one value in the *vegetable* frame. For both categories, using only category properties makes a mean difference of +.1. Using cleaned data makes a mean difference compared to all data of +.08 for *fruit* and +.13 for *vegetable*. The differences between using attribute diagnosticity or none is small for both categories (-.02 for *fruit*, +.03 for *vegetable*), but the mean difference between property level and no diagnosticity is +.08 for *fruit* and +.06 for *vegetable*. In both cases, the highest correlations (*fruit*: .84-.87,

vegetable: .67-.69) are found when using category properties only and cleaned data, regardless of where the diagnosticity weights are applied. Without cleaning, the models with diagnosticity on the property level and without diagnosticity have higher correlations (.82 and.85) than the model with attribute diagnosticity (.67) for *fruit*, for *vegetable* these models predict in the same range (.53-.55). The best model when contrast properties are included is again the property model with only slightly lower correlations between .79 and .81, next the attribute model (.60 and .78) and last the one without diagnosticity (.63 and .70) for *fruit*.

Table 62: Results for all predictions compared for a) fruit and b) vegetable.

a)

diag	properties	datatype	*r*	p	SE	CI
attribute	category	all	.67	.001	.144	[0.39, 0.96]
		cleaned	.84	.000	.050	[0.75, 0.94]
	all	all	.60	.003	.182	[0.25, 0.96]
		cleaned	.78	.000	.060	[0.66, 0.89]
property	category	all	.85	.000	.064	[0.73, 0.97]
		cleaned	.87	.000	.040	[0.79, 0.95]
	all	all	.79	.000	.087	[0.62, 0.96]
		cleaned	.81	.000	.045	[0.72, 0.9]
none	category	all	.82	.000	.081	[0.66, 0.98]
		cleaned	.83	.000	.042	[0.75, 0.91]
	all	all	.63	.002	.141	[0.35, 0.9]
		cleaned	.70	.000	.069	[0.57, 0.84]

attribute - property	-.11
attribute - none	-.02
property - none	.08
category - all	.10
all - cleaned	-.08

b)

diag	properties	datatype	*r*	p	SE	CI
attribute	category	all	.51	.015	.165	[0.19, 0.83]
		cleaned	.67	.001	.145	[0.38, 0.95]
	all	all	.45	.037	.205	[0.05, 0.85]
		cleaned	.59	.004	.162	[0.27, 0.9]
property	category	all	.53	.008	.157	[0.22, 0.84]
		cleaned	.68	.000	.144	[0.39, 0.96]
	all	all	.51	.012	.168	[0.18, 0.84]
		cleaned	.62	.002	.166	[0.3, 0.95]
none	category	all	.54	.007	.153	[0.24, 0.84]
		cleaned	.69	.000	.154	[0.38, 0.99]
	all	all	.38	.075	.209	[-0.03, 0.78]
		cleaned	.48	.019	.196	[0.1, 0.87]

attribute - property	-.03
attribute - none	.03
property - none	.06
category - all	.10
all - cleaned	-.13

7.1.4 **Parameter-fitting**

To find the optimal solutions for all models, the Microsoft Excel add-in program Solver was used to maximise correlations and with the constraints that the sum-product of the diagnosticity values with the category probabilities is 1 and that each diagnosticity value is smaller or equal 1 and for attribute models that the diagnosticity per attribute should be the same. The objective variable was the correlation between rated and predicted variable. The variable cells to be changed were the diagnosticities. The solver method GRG (generalised reduced gradient) was used.

Parameter-fitting is a post-facto method that is no prediction. It was used to analyse the optimal solutions and possibly identify weaknesses in the model. To become a prediction, the method of cross-validation would have to be used in which the parameters are fitted with one half of the data and then used to predict the other half of the data. This is out of the scope of this thesis.

The results for *fruit* are in Table 63 for the models with category properties only. The parameter fitting leads to almost perfect correlations between .89 and .95. In all optimisations, the diagnosticity of the attribute MAIN-NUTRITIONAL-COMPONENT was raised and the ones of SEEDS and TASTE reduced or set to 0. This is in line with the observation in section 6.4.5 that the mean probability rating for the property *has sugar* has a very high correlation (.87) with the typicality ratings.

Table 63: Results of parameter-fitting diagnosticity for fruit with category properties only.

property	attribute	value	diag P	diag A	diag A all	diag A cleaned	diag P all	diag P cleaned
will be made into juice	CONSUMED-IN	Juice	.11	.26	-	-	-	-
will be eaten as dessert	CONSUMED-IN	Dessert	.22	.26	-	-	.19	.30
grew on a tree	HOW-GROWN	Trees	.12	.15	.08	.00	.06	.02
grew on a plant	HOW-GROWN	On Plants	.06	.15	.08	.00	-	.27
is juicy	JUICINESS	Juicy	.15	.18	.16	.05	.16	.03
has sugar	MAIN-NUTRITIONAL-COMPONENT	Sugar	.13	.15	.86	.98	.65	.42
has pips/seeds	SEEDS	Yes	.12	.14	-	.01	-	.06
is sweet	TASTE	Sweet	.15	.18	-	.03	-	.03
tastes tart	TASTE	Tart	.14	.18	-	.03	.08	.07
		optimised correlation			.89	.93	.90	.95

Figure 78 shows the typicality contributions of the models with uncleaned data for weights on the attribute level (ca) and weights on the property level (cp) compared to their fitted solutions. Raising the influence of the MAIN-NUTRITIONAL-COMPONENT attribute leads to very good predictions and even corrects the predictions for *banana*, which is predicted too low in all models. However, except for the attributes JUICINESS and CONSUMED-IN-Dessert for cp, the contribution of all other attributes becomes very small.

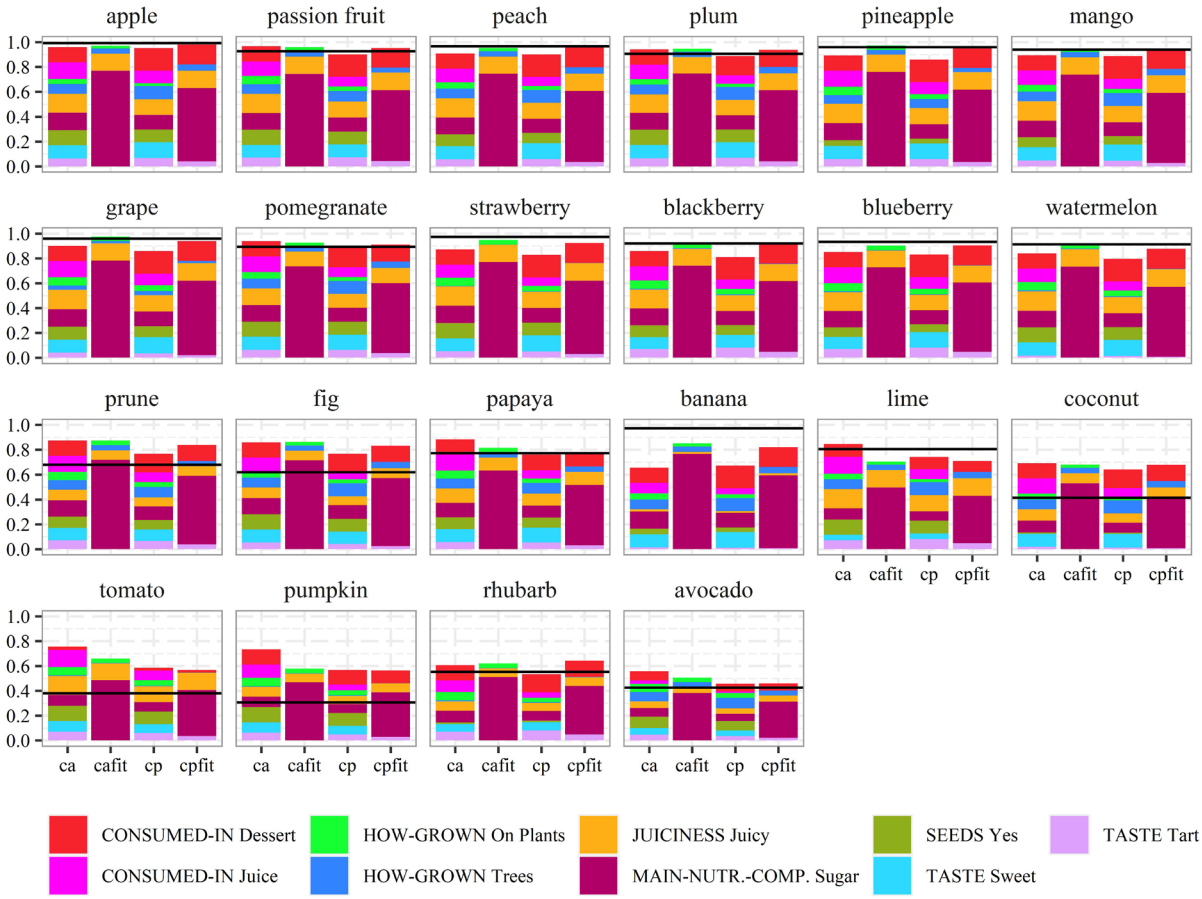

Figure 78: Typicality contributions for predictions with category properties and diagnosticity applied to properties (cp) or attributes (ca) compared to fitted data for fruit.

Figure 79 compares the cleaned versions of the models, cac and cpc, with their fitted versions. Again, MAIN-NUTRITIONAL-COMPONENT has a very high typicality contribution. Whenever it was excluded in cleaning because of low intersubjective agreement, a combination of *is juicy*, *is made into juice*, *tastes sweet* and *is eaten as dessert* compensates for it, both in the fitted and unfitted versions. The difference between the predictions of the cleaned models and their fitted versions is minimal.

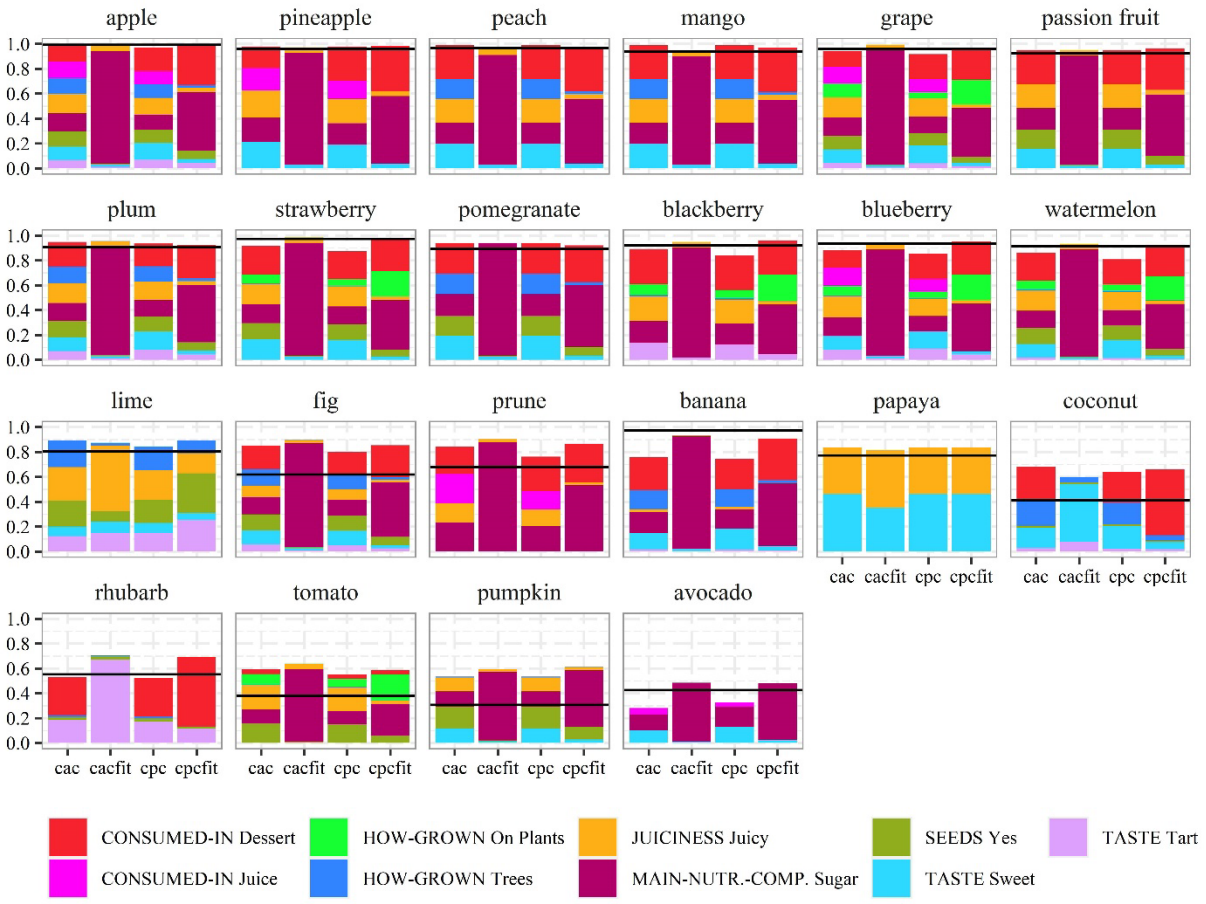

Figure 79: Typicality contributions for predictions with cleaned data, category properties and diagnosticity applied to properties (cpc) or attributes (cac) compared to fitted data for fruit.

Table 64 shows the results for the models that include contrast properties. Again, the fitting leads to almost perfect correlations between .89 and .95. In all optimisations, the diagnosticities of the attributes ROOTS and USED-IN-COOKING are reduced to 0 and the one for MAIN-NUTRITIONAL-COMPONENT is raised. In the models with uncleaned data, the contrast property *has a heart* receives a very high weight, but in the models with cleaned data it receives a weight of 0. For the attribute model without cleaning, many properties received a small or zero weight. *Will be eaten as dessert* receives a high weight in the models with diagnosticity weights on the property level.

Table 64: Results of parameter-fitting diagnosticity for fruit with all properties.

property	attribute	value	diag P	diag A	diag A all	diag A cleaned	diag P all	diag P cleaned
will be eaten as dessert	CONSUMED-IN	Dessert	.14	.18	.03	-	.16	.35
will be made into juice	CONSUMED-IN	Juice	.11	.18	.03	-	-	-
will be eaten in soup	CONSUMED-IN	Soup	.02	.18	.03	-	-	-
has a heart	HEART	Yes	.06	.07	.54	-	.76	.03
grew on a tree	HOW-GROWN	Trees	.12	.15	.05	.00	.05	.04
grew on a plant	HOW-GROWN	On Plants	.10	.15	.05	.00	-	.45
grew in the ground	HOW-GROWN	In the Ground	.04	.15	.05	.00	-	-
is juicy	JUICINESS	Juicy	.12	.16	.13	.05	.12	-
has sugar	MAIN-NUTRITIONAL-COMPONENT	Sugar	.12	.15	.77	.98	.58	.24
will be boiled	PREPARATION-METHOD	Boiled	.04	.05	-	-	-	-
will be roasted	PREPARATION-METHOD	Roasted	.03	.05	-	-	-	.05
has roots	ROOTS	Yes	.04	.04	-	-	-	-
has pips/seeds	SEEDS	Yes	.11	.14	-	.01	-	.05
is sweet	TASTE	Sweet	.12	.16	-	.03	-	.01
tastes tart	TASTE	Tart	.12	.16	-	.03	.09	.09
will be used in cooking	USED-IN-COOKING	Yes	.06	.08	-	-	-	-
		optimised correlation			.89	.93	.91	.95

All in all, the optimisations show that raising the influence of the properties *has sugar* and *is juicy* improves the correlations, in some cases significantly. The fact that excluding or strongly reducing the influence of the vegetable properties in the unified models raises the correlation speaks for the superiority of using category properties only.

Table 65 shows the results for the models with category properties for vegetable. The optimisations improved all correlations, between +.15 and +.22, to values between .66 and .8. The model with cleaned data sets all weights except for *will be eaten in soup* to very small values or 0. The models with uncleaned data set the highest weight to *has a heart* and *will be eaten in soup* and either *will be used in cooking* for the attribute version or *will be boiled* in the property version. Only *will be roasted* receives consistently a low or 0 weight.

Table 65:Results of parameter-fitting diagnosticity for separate models for vegetables.

property	attribute	value	diag P	diag A	diag A all	diag A cleaned	diag P all	diag P cleaned
will be eaten in soup	CONSUMED-IN	Soup	.26	.30	.36	1.00	.33	1.00
has a heart	HEART	Yes	.10	.11	1.00	.10	1.00	.04
grew in the ground	HOW-GROWN	In the Ground	.20	.22	-	.03	-	.05
will be roasted	PREPARATION-METHOD	Roasted	.23	.26	.05	.01	-	-
will be boiled	PREPARATION-METHOD	Boiled	.19	.26	.05	.01	.36	.06
has roots	ROOTS	Yes	.20	.23	.11	.04	.13	-
will be used in cooking	USED-IN-COOKING	Yes	.07	.08	.24	.02	-	.01
		optimised correlation			.66	.74	.75	.80

Figure 80 compares the typicality contributions for the attribute (ca) and property (cp) version of the uncleaned data. The fitted versions predict the typicality of the SCs on the high typicality level less good, but the predictions on the lower end of typicality fit much better. As we identified few highly typical SCs, this raises the correlation. Raising the weight for *has a heart* in the versions with uncleaned data only has a consequence for the typicality prediction of *lettuce*, the only SC in the whole dataset that has a mean rated probability greater than 0 for this property.

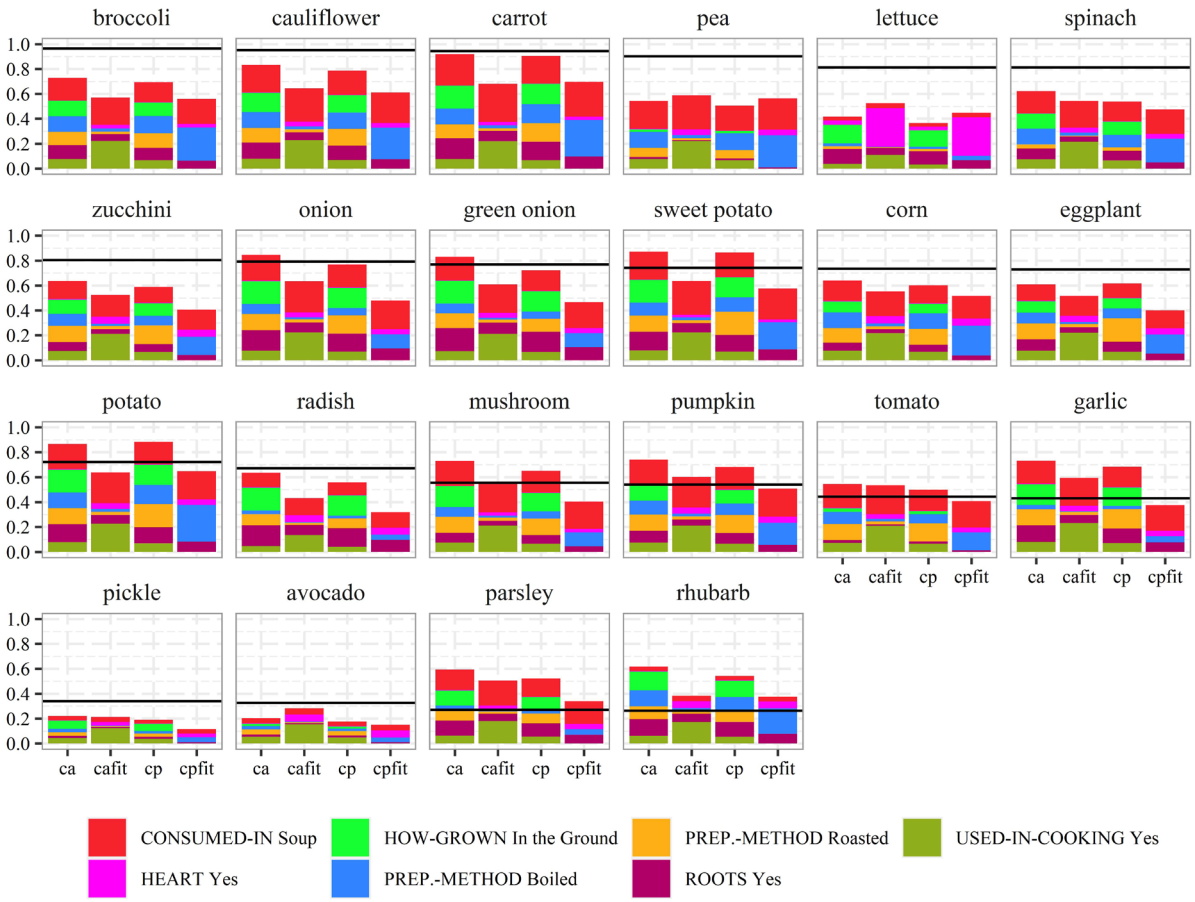

Figure 80: Typicality contributions for predictions with category properties and diagnosticity applied to properties (cp) or attributes (ca) compared to fitted data for vegetables.

Figure 81 shows the typicality contribution for the cleaned data. Both fitted version have a high weight on *will be eaten in soup*. Wherever this property is excluded in cleaning, *will be boiled*, *will be used in cooking*, *grew in the ground* and *will be roasted* are the main contributions. Like in the cleaned models without fitted parameters, the high typicality of *broccoli, carrot* and *cauliflower* and in addition *pea* is well predicted in the fitted versions, while *lettuce* and *spinach* are predicted too low regardless of which model is used.

Figure 81: Typicality contributions for predictions with cleaned data, category properties and diagnosticity applied to properties (cpc) or attributes (cac) compared to fitted data for vegetables.

Table 66 shows the results for the models with contrast properties for *vegetable*. *Has a heart* receives again a high diagnosticity weight in the models with uncleaned data, which again gives *lettuce* a high typicality contribution. For the highest correlation (.82), which is found for the model with diagnosticity weights on the property level and cleaned data, *will be eaten in soup* receives a very high weight and the other weights are at or below .09. This leads to the typicality contributions coming mainly from *will be eaten in soup* and when this property was removed in cleaning, *grew in the ground*, *will be boiled* and *has sugar* compensate for it. *Has sugar* receives a high weight in both attribute models, as well as *will be used in cooking*.

Table 66: Results of parameter-fitting diagnosticity for unified models for vegetables with difference to original diagnosticity in brackets.

property	attribute	value	diag P	diag A	diag A all	diag A cleaned	diag P all	diag P cleaned
will be eaten in soup	CONSUMED-IN	Soup	.17	.20	.08	.07	.28	.92
will be made into juice	CONSUMED-IN	Juice	.06	.20	.08	.07	-	.04
will be eaten as dessert	CONSUMED-IN	Dessert	.02	.20	.08	.07	-	-
has a heart	HEART	Yes	.12	.15	.89	.45	1.00	-
grew in the ground	HOW-GROWN	In the Ground	.15	.18	.03	.06	.11	.08
grew on a plant	HOW-GROWN	On Plants	.08	.18	.03	.06	.09	-
grew on a tree	HOW-GROWN	Trees	.05	.18	.03	.06	.19	.00
is juicy	JUICINESS	Juicy	.04	.05	-	-	.03	-
has sugar	MAIN-NUTRITIONAL-COMPONENT	Sugar	.05	.06	.40	1.00	.09	.04
will be roasted	PREPARATION-METHOD	Roasted	.16	.19	-	-	-	-
will be boiled	PREPARATION-METHOD	Boiled	.15	.19	-	-	.23	.09
has roots	ROOTS	Yes	.15	.18	.08	.10	-	-
has pips/seeds	SEEDS	Yes	.06	.07	-	-	-	-
tastes tart	TASTE	Tart	.05	.06	-	-	-	-
is sweet	TASTE	Sweet	.04	.06	-	-	-	.02
will be used in cooking	USED-IN-COOKING	Yes	.12	.14	.39	.14	-	.00
		optimised correlation			.67	.72	.76	.82

7.2 Individual predictions

So far, our predictions of typicalities from probabilities used the mean ratings of the participants. Another interesting question is how well each participant's property probability ratings predict their own typicality ratings and the mean typicality ratings when used as input for the models. All models described above were computed for each participant and correlated with their individual typicality rating as well as with the mean typicality rating.

The mean correlations of all participants are in Table 67 for *fruit*. The correlations of the individual predictions with the individual typicality ratings are lower than the correlations with the mean typicality ratings with an average difference of -.1 and a higher SD with an average difference of +.11. Compared to the predictions with mean probabilities presented in the last section, the difference between applying diagnosticity weights on the attribute vs. property level is lower (-.11 vs. -.03) and the attribute models have slightly higher correlations than the models with no diagnosticity weights (+.03 for correlations with own ratings and +.05 with mean ratings). The models with the highest correlations between .65 and .68 are the models with category properties and cleaned data, the diagnosticity application level makes no difference in these cases (the maximum difference is .03). The use of cleaned data makes a higher difference for the correlations with the mean rating (+.08) than for the correlations with the own rating (+.03).

Table 67: Mean correlation of model predictions using participants' probability ratings with mean typicality rating and participants' typicality ratings, and mean correlation, differences for means and SDs for fruit.

diag	properties	datatype	mean correlation with own rating	SD	mean correlation with mean rating	SD	diff mean	diff SD	correlation with mean probability
attribute	category	all	.48	.27	.57	.15	-.10	.12	.67
		cleaned	.52	.30	.68	.13	-.16	.17	.84
	all	all	.42	.23	.48	.17	-.06	.06	.60
		cleaned	.47	.26	.61	.15	-.13	.11	.78
property	category	all	.52	.29	.63	.14	-.11	.14	.85
		cleaned	.53	.30	.67	.13	-.14	.17	.87
	all	all	.47	.26	.55	.16	-.08	.10	.79
		cleaned	.49	.27	.62	.14	-.12	.13	.81
none	category	all	.49	.28	.59	.15	-.10	.13	.82
		cleaned	.51	.29	.65	.13	-.14	.16	.83
	all	all	.36	.21	.39	.18	-.03	.02	.63
		cleaned	.39	.23	.49	.18	-.09	.05	.70
attribute - property			-.03		-.03				-.11
attribute - none			.03		.05				-.02
property - none			.06		.09				.08
category - all			.07		.11				.10
all - cleaned			-.03		-.08				-.08
own rating - mean			-.10						

Figure 82 shows the individual predictions split by models. There are some participants for whom the predictions are higher than the mean and some for which they are significantly lower, with values ranging from -.64 to .87. Due to this high variance, the mean does not reflect the central tendency of the data well.

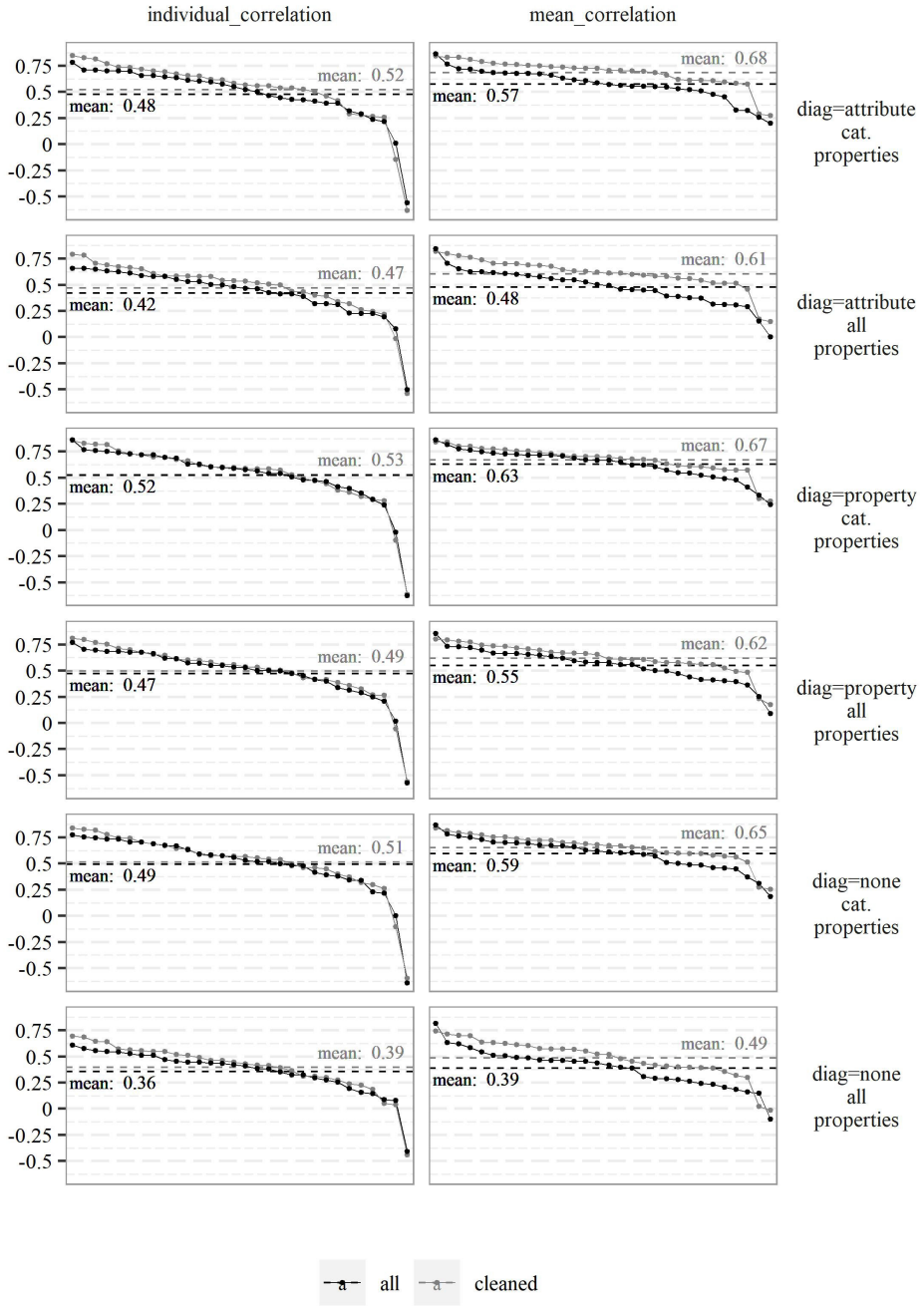

Figure 82: Individual predictions of each participant ordered by size of correlation for fruit per model with means (dashed lines) for all data (black) and cleaned data (grey).

For *vegetable*, the correlation between rated typicality and mean predictions with individual ratings are in Table 68. The mean correlation of individual predictions with individual typicality ratings are between .21 and .36. The highest correlations are found for the cleaned version with category properties only and diagnosticity on the property level (.35) or no diagnosticity weights (.36). Using cleaned data increased the correlations by .09 on average. All other differences between the models change the correlations by less than |.05|. The mean correlation of individual predictions with mean typicality ratings are between .13 and .54. Again, the highest are found for the cleaned versions with category properties only and diagnosticity on the property level (.51) or no diagnosticity weights at all (.54). For the correlations with the mean ratings, it makes a high mean difference (+.16) to use cleaned data and to use category

properties only (+.11). Also, the correlations are higher with diagnosticity weights on the property level than with no diagnosticity weights with a mean difference of +.07. This difference comes solely from the versions in which contrast properties are included, where property diagnosticity has much higher correlations than no diagnosticity.

Table 68: Mean correlation of model predictions using participants' probability ratings with mean typicality rating and participants' typicality ratings and mean correlation, differences for means and SDs for vegetable.

diag	properties	datatype	mean correlation with own rating	SD	mean correlation with mean rating	SD	diff mean	diff SD	correlation with mean probability
attribute	category	all	.24	.23	.35	.15	-.12	.09	.53
		cleaned	.33	.26	.49	.15	-.16	.11	.68
	all	all	.21	.24	.27	.17	-.05	.07	.46
		cleaned	.31	.22	.43	.14	-.12	.08	.60
property	category	all	.25	.23	.37	.13	-.12	.10	.55
		cleaned	.35	.24	.51	.15	-.16	.09	.69
	all	all	.23	.23	.32	.15	-.09	.08	.53
		cleaned	.33	.23	.47	.14	-.14	.09	.63
none	category	all	.25	.24	.38	.13	-.13	.11	.56
		cleaned	.36	.24	.54	.15	-.17	.09	.70
	all	all	.23	.25	.13	.20	.10	.05	.39
		cleaned	.27	.20	.35	.12	-.09	.07	.49
attribute - property			-.02		-.03				-.03
attribute - none			-.01		.04				.03
property - none			.01		.07				.07
category - all			.03		.11				.10
all - cleaned			-.09		-.16				-.13
own rating - mean			-.11						

Figure 83 shows the individual predictions per model. Compared to fruit, the range is slightly less extreme and the values range between -.32 and .84. The correlations with individual typicality ratings vary strongly between participants and have a high SD.

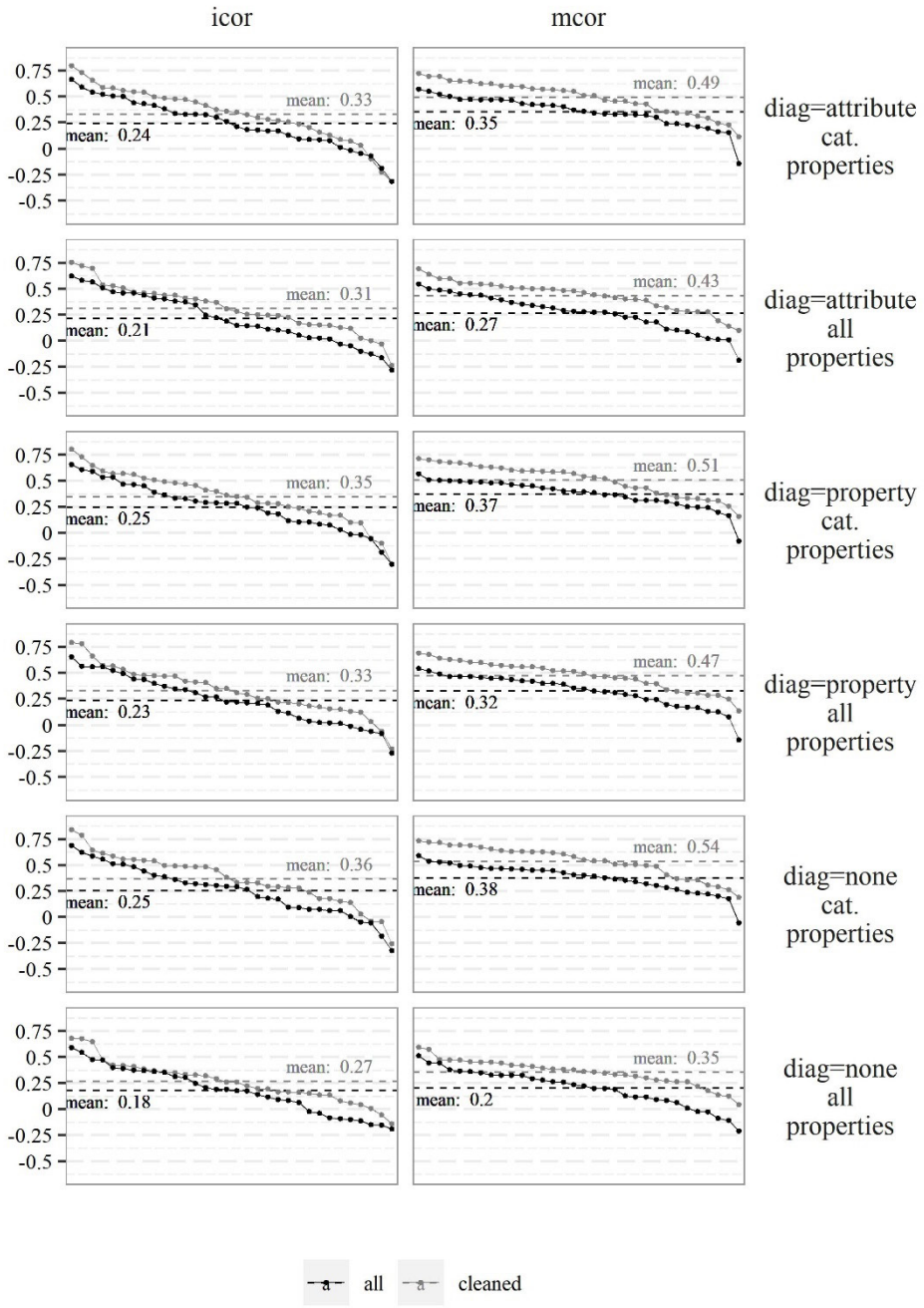

Figure 83: Individual predictions for vegetable per model with means (dashed lines) for all data (black) and cleaned data (grey) for individual typicality ratings (icor) and mean typicality ratings (mcor).

This analysis shows that the prediction of typicality with property probability ratings works very well for some participants, but not for all. Several explanations for this can be found. First, the predictions are a lot more fine-grained than the typicality ratings, because the former are based on a 100-point-scale and the latter on a 7-point-scale. Perfect Pearson correlations can therefore not be expected. Secondly, many participants seem to have understood the ratings as possibility instead of frequency ratings, which might lead to lower correlations due to inaccurate input.

7.3 Discussion

Typicality prediction with probabilistic prototype frames was a success. It is not surprising that using cleaned data improves the correlations substantially, due to the low intersubjective agreement found for one third resp. half of the SC-property-pairs as discussed in section 6.3.5. Predictions with category properties have higher correlations than predictions that include contrast category properties. The fact that the correlations for fruit rise considerably by +.18 and are more stable in jackknifing if the diagnosticity weights are applied on the property level indicates that using the maximum per attribute might not be the best choice and that one very diagnostic value in an attribute does not raise the importance of all values of that attribute. However, as the attributes were only complemented with a low number of values, this conclusion is tentative and should be confirmed by testing the two models with frames that are complemented with more attribute values (see section 7.4). The application level of diagnosticity makes no difference for the cleaned models with correlations between .83 and .87 for *fruit* and between .68 and .79 for *vegetable*, so it might be the case that diagnosticity is helpful for identifying relevant category properties, but not an essential part of the formula. It does however make a difference when contrast category properties are incorporated, with .08 – .11 higher correlations for *fruit* and .11 – .14 higher correlations for *vegetable*, where it leads to a higher typicality contribution of category properties.

While the correlations are made almost perfect with parameter-fitting for *fruit*, *vegetable* reaches lower values even here. The properties that were selected are not able to account for the high typicality of *lettuce* and *spinach*, as one of them is never cooked and both of them are not eaten in soup, which are the properties that contribute most to typicality in the frames I identified. They also grow on the ground and this growing condition was not included. I suspect that adding the properties *will be eaten raw* and *grows on the ground* would improve the typicality predictions for *vegetable*. That the correlations for *vegetable* are in general lower than those for *fruit* might also in part be explained by the finding that the typicality orderings have smaller correlations between American and British studies (see section 3.2.6) and one third of my participants were American.

In addition to the correlations, two criteria are available to evaluate the success of the predictions: first, how well they fare compared to the correlations reported in the literature and, second, how well they fare compared to the correlation of each property's mean probability rating with the typicality data.

Rosch and Mervis (1975) report corelations with mean rated typicality of .85 for fruit and .84 for vegetables and Smith et al. (1988) report .75 for fruit and .40 for vegetables. The predictions with cleaned data are in the same range as the correlations from Rosch and Mervis for *fruit* and both predictions are above the correlations from Smith et al. for both *fruit* and *vegetable*, except for the *fruit* predictions without cleaning for which they are slightly lower.

Table 69 shows the correlations of each property's mean probability rating conditional on a subcategory with mean rated typicality of this subcategory (that has already been presented in section 6.4.5). For *fruit*, the properties *has sugar* and *will be eaten as dessert* have high correlations of .87 resp. .78 and predict typicality slightly better resp. only slightly worse than the probabilistic prototypes with cleaned data and better than the model with all data. For *vegetable*, *will be boiled* is correlated (.59) with typicality only slightly lower than the models with cleaned data and higher than the models with all data.

Table 69: Correlation between each property's mean probability rating with mean typicality for fruit and vegetables.

		fruit		vegetable	
	property	r	p	r	p
fruit properties	has sugar	.87	.000	-.05	.836
	will be eaten as dessert	.78	.000	-.39	.075
	is sweet	.73	.000	-.02	.946
	will be made into juice	.59	.004	-.14	.528
	is juicy	.52	.013	-.20	.362
	grew on a tree	.19	.402	-.35	.105
	has pips/seeds	.15	.514	-.38	.080
	tastes tart	.08	.739	-.57	.005
	grew on a plant	.02	.922	-.04	.867
vegetable properties	will be boiled	-.44	.004	.57	.005
	will be eaten in soup	-.72	.013	.52	.013
	will be used in cooking	-.66	.054	.41	.057
	grew in the ground	-.43	.143	.30	.172
	has roots	-.47	.152	.29	.183
	will be roasted	-.58	.288	.22	.317
	has a heart	-.38	.625	.11	.623

What needs to be taken into consideration however is that the explanatory power of probabilistic prototypes exceeds the one of the correlation with a single property considerably. Clearly, *fruit* and *vegetable* SCs are connected by more properties than one or two and therefore, even with slightly lower correlations, they should be preferred, because they explain more.

7.4 **Possible modifications**

While the concept *fruit* is well represented by the attribute-value-structure that was chosen, the same does not seem to be true for the concept *vegetable*. Properties that characterise very typical vegetables seem to be missing, as the predictions are consistently too low for the highly typical SCs *lettuce* and *spinach*.

Furthermore, the superior role of diagnosticity weights on the property level should be confirmed by using attributes that are complemented with additional values. Considering that the embedding of contrast properties decreased the correlations for all models, an alternative design in which different properties are tested for each of the two categories is preferable. This would mean that each category representation could be complemented with more attribute values without prolonging the experiment. Based on the data from experiment 1 and 2, prototype frames that incorporate these changes are in Figure 84 for *fruit* and Figure 85 for *vegetable*.

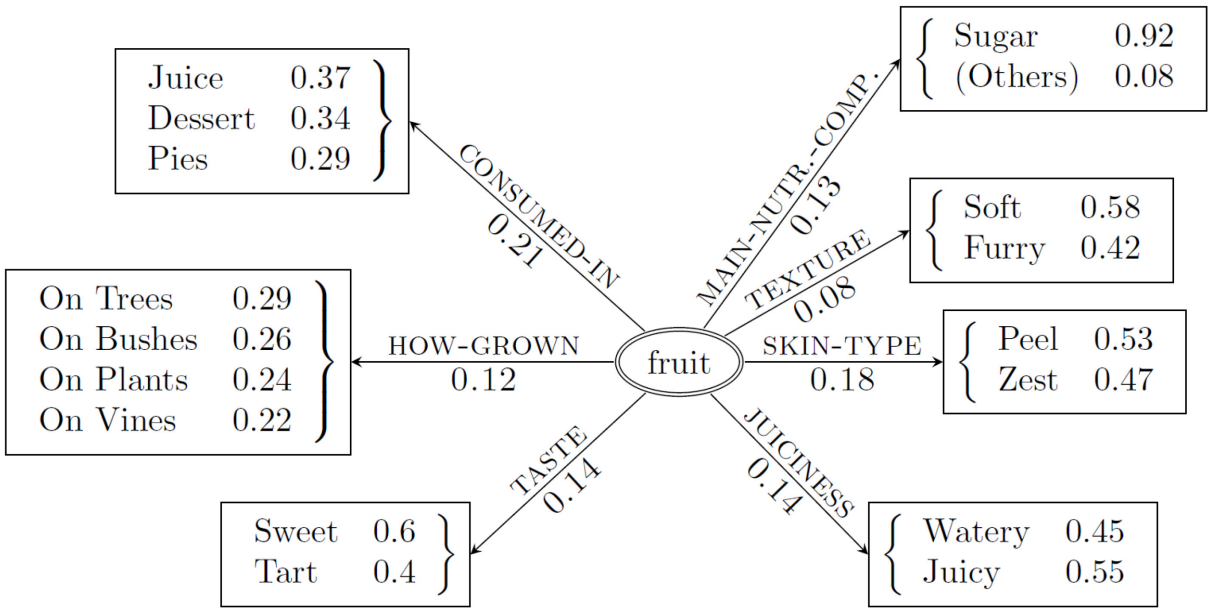

Figure 84: Alternative prototype frame for fruit with data from experiment 1 and 2.

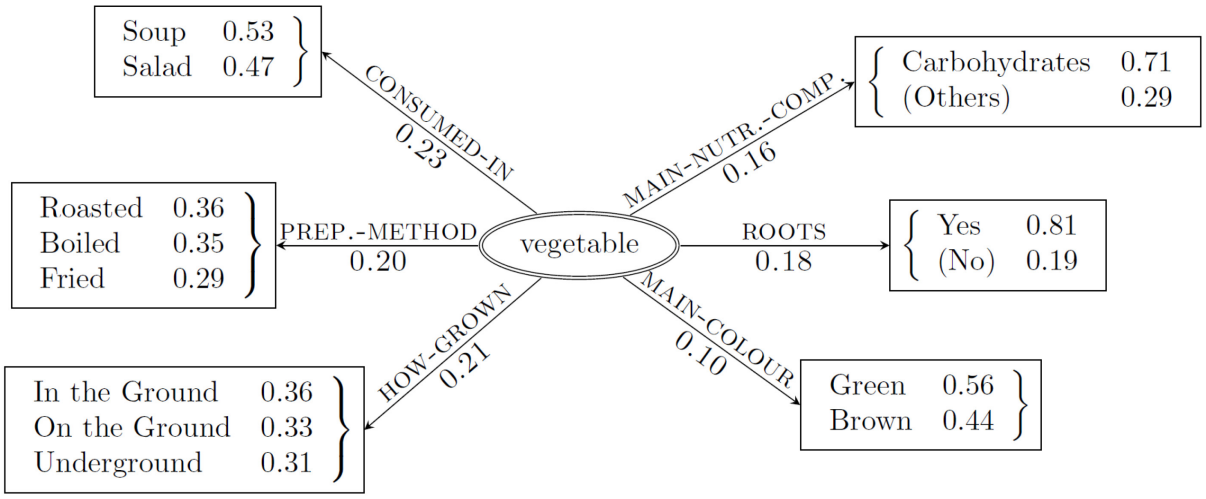

Figure 85: Alternative prototype frame for vegetables with data from experiment 1 and 2.

In the context of the typicality meta-analysis, it was found that the *vegetable* typicality orderings only had high correlations in similar geographic regions. In particular, the correlations between typicality orderings from the USA and Great Britain were low. It would therefore be advisable to only recruit participants from either region.

In section 3.2, I showed that typicality ratings are not as intersubjectively stable as is commonly assumed, in particular on the medium typicality level. Therefore, typicality as such might not be the best variable to confirm the cognitive plausibility of prototype representations. Instead, speed of categorisation of the same SCs could be a better candidate. It measures indirectly how straightforwardly participants assign the category term.

Or, as I discussed in section 3.4, mean typicality ratings should be interpreted less stringent. Instead of each value reflecting a distinct level, there might only be three distinguishable typicality levels with specific characteristics. A model could be regarded as cognitively plausible if it correctly predicts high, medium and low typicality.

8 Predicting typicality with family resemblance and contrast

This chapter analyses two influential models of typicality prediction that were proposed in the past: the family resemblance score proposed in Rosch and Mervis (1975) (section 8.1) and the contrast model introduced in Smith et al. (1988) (section 8.2). The chapter has two purposes: first, the models are tested for generalisability by applying them to new datasets from the literature and the second step is to test how well the probability judgments gathered for this thesis predict typicality when used as an alternative input for the models.

8.1 Family resemblance score

Rosch and Mervis (1975) show that the typicality of SCs for Cs is strongly correlated with their family resemblance score (FRS). An example calculation of their model was already presented in Figure 30 and is repeated as Figure 86. They characterise Cs with properties P_k and SCs with binary application scores (AS) for these properties. The FRS of a property is the sum of its ASs over all SCs. The FRS of an SC is determined as the weighted sum of its ASs (assuming there are n SCs and m properties of SC_i):

$$\text{(FRS)}\quad FRS(SC_i, C) = \sum_{k=1}^{m} FRS(P_k) \cdot AS_{P_k}(SC_i),$$

$$\text{with } FRS(P_k) = \sum_{i=1}^{n} AS_{P_k}(SC_i) \text{ and } AS_{P_k}(SC_i) \in \{0,1\}.$$

In words: the family resemblance score of a subcategory SC_i in a category C is obtained by counting, for each property P_k of the SC_i, how many subcategories of the category have this property, and then summing up these counts for all properties possessed by SC_i. To find relevant P_k, they collected property lists for 120 SCs of 6 Cs by asking 20 participants per SC to list its characteristic properties in 90 seconds. Each participant listed properties for one SC of each C. The property application score for each SC, $AS_{P_k}(SC_i)$, was determined by two judges who deleted "obviously wrong" properties and added those properties "obviously true" for other SCs. The property FRS ranged between 1 to 20 depending on how many SCs were rated to have P_k. The SCs' FRSs were then correlated with their mean typicality ratings. They also calculated correlations resulting from applying the natural logarithm to the properties' FRS (FRS_{ln}), which reflects the thought that properties shared by more SCs should receive a higher relative weight than properties which are shared only by few SCs.

property	apple	blueberry	pear
is sweet	1 · 3	1 · 3	1 · 3
is round	1 · 2	1 · 2	0 · 2
is blue	0 · 1	1 · 1	0 · 1
is green	1 · 2	0 · 2	1 · 2
	3+2+0+2=7	3+2+1+0=6	3+0+0+2=5

property	FR(P)
is sweet	3
is round	2
is blue	1
is green	2

Figure 86: Example calculation of the family resemblance score for three hypothetical SCs and four hypothetical properties.

Rosch and Mervis found that SCs with high mean typicality ratings in Rosch (1975b) also shared many properties with other SCs, while SCs that had been rated rather untypical shared only few or no properties. Furthermore, they found a high rank order correlation between FRS

and typicality. The reported results are in Table 70. The correlations are high (>.8) for all investigated categories and there is no remarkable difference for the logarithmic FRS. Furthermore, the FRS is highly correlated with the SCs' distance from the origin in a multidimensional scaling solution for five categories investigated in the connection of a larger study.

Table 70: Reported results from Rosch, Mervis (1975): Spearman rank order correlations of typicality with family resemblance score (FR) and logarithmic FRS (FRln) as well as FRS with distance from origin in a multidimensional scaling (MDS) experiment.

	FR-typ	FRln-typ	FR-MDS
furniture	0.88	0.84	0.89
vehicle	0.92	0.9	0.94
weapon	0.94	0.93	0.95
fruit	0.85	0.88	0.92
vegetable	0.84	0.86	0.9
clothing	0.91	0.88	

A data type that is very similar to theirs was made available by De Deyne et al. (2008): "exemplar by feature applicability matrices" (in our terminology: SC by property applicability matrices) and I used them to replicate their results (section 8.1.1). Then, I used the property probability data I collected to calculate the FRS by using different probability thresholds as a criterion for applicability instead of ratings by judges (section 8.1.2).

8.1.1 **De Deyne et al. data**

From De Deyne et al. (2008), two kinds of SC by property applicability matrices are available: ones that include properties generated for SCs and ones that include properties generated for Cs. Both include applicability judgements by 4 participants for 30 SCs. All properties mentioned for at least 4 out of 30 SCs in a property generation task with no time limit were used for the SC matrices. Properties generated for the Cs with a productive frequency (PF) of at least 2 were used for the C matrices. For *fruit* SCs, 5,701 properties were generated, 741 of which were unique and 233 of those were generated for at least 4 SCs[34]. For *fruit* as C, 205 properties were generated, 52 were unique and 32 were generated by at least 2 participants. For *vegetable* SCs, 5,741 properties were generated, 903 unique and 291 for at least 4 SCs[35]. For *vegetable* as C, 193 properties were generated, 50 distinct and 30 of them have a PF of at least 2. While the SC property applicability matrices have the same structure as the ones from Rosch and Mervis (1975), the C property applicability matrices are interesting with regard to the question whether the prototype is better generated from SC properties or from C properties (see section 4.2 for a comparison).

[34] For fruit SCs, the Dutch original contains "behaard" and "harig" which are both translated to "hairy" in the English translation provided by the authors. After some research, "harig" was translated to "shaggy" to keep 233 unique properties.

[35] For vegetable SCs, the Dutch original contains "word gekweekt" and "word geteeld" which are both translated to "is cultivated" in English. After some research, "word gekweekt" was translated to "is grown" to keep 291 unique properties.

Different from the original procedure which employed judges to determine applicability, De Deyne et al. provide data on how many participants out of 4 agreed on the applicability of the properties for the different SCs. The distribution of the applicability ratings is in Figure 87. All four participants agreed for less than 20% of the properties in all four datasets. Except for *fruit* category properties, the number of properties for which only one participant saw applicability is roughly the same as the one for which there was complete agreement. For roughly 50% of the properties from each property list, the participants agreed on the non-applicability of properties, i.e., those that have an application score of 0.

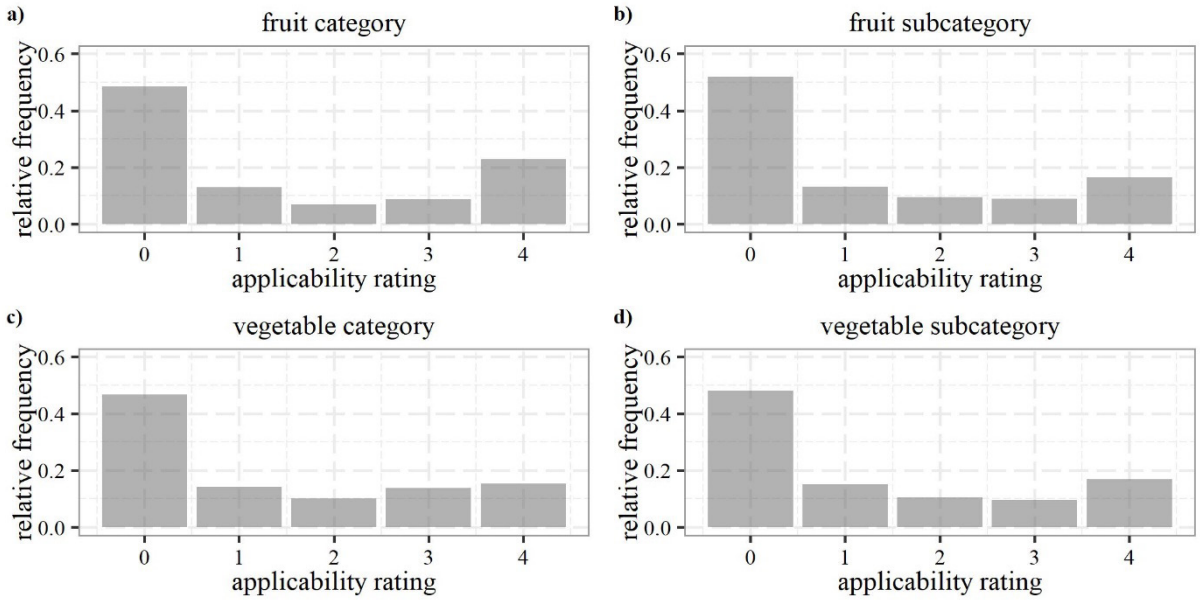

Figure 87: Frequency distribution of applicability ratings in De Deyne et al. (2008) data.

The FRS was calculated in five different ways to see the difference between different applicability thresholds. In the strictest version, only those SC-property-pairs for which all 4 participants agreed on applicability were included in the calculation (AS4). The others required only agreement from 3 (AS3), 2 (AS2) or 1 (AS1) participant. The fifth version is the weighted application score (WAS) score proposed in Djalal et al. (2017). It uses the sum of applicability judgements per property over all SCs as $FR(P_k)$ and thus assumes applicability to be graded (see section 4.3). Example calculations for all interpretations of applicability are in Table 71 for 12 of the 30 SCs they used and 4 properties. The WAS has higher values in general because it sums up the applicability judgments, while the standard FRS treats applicability as binary which results in the maximal value being determined by the number of SCs that were included in the experiment (12 in the example). This means that properties with high agreement receive a higher weight in the WAS score. For example, the property *as a snack*, which applies to almost all SCs according to almost all participants, contributes to typicality almost double (43) than the next most applicable property *a little sourish* (28). In the standard FRS, the differences between property FRSs are much smaller, for example 12 for *as a snack* and 9 for *a little sourish* for the AS1. The sum of all applicable properties' FRSs is then the SCs' FRS. For the WAS, each property is weighted again with the number of positive applicability judgments and then summed. The AS1 introduces only 3 typicality levels in the example: 29, 21 and 20. AS2 has 5, AS3 and AS4 have 4 and WAS has 10 and thus predicts the most fine-grained typicality ordering. The difference in predicted typicality levels is smaller for the calculations that use all properties and SCs. For the *fruit* category, there were between 24 and 26 unique scores. For *fruit* SCs, there were between 27 and 30. For *vegetable* with category properties there were

between 20 and 29 and for *vegetable* SCs there were 29 unique scores for all applicability criteria. The WAS produced 30 unique scores in every case.

Table 71: Example calculations of different FRS versions for properties and SCs for De Deyne et al. (2008) fruit data.

property	red currant	strawberry	apricot	pineapple	apple	banana	blueberry	grapefruit	pumpkin	plum	orange	fig	FR(P) AS=1	FR(P) AS=2	FR(P) AS=3	FR(P) AS=4	FR(P) AS=WAS
afrika	0	0	0	2	1	4	0	1	1	1	1	2	8	3	1	1	13
as a snack	4	3	4	4	4	4	4	4	1	4	4	3	12	11	11	9	43
monkeys like to eat it	0	0	0	0	0	4	0	0	0	0	0	0	1	1	1	1	4
a little sourish	4	2	1	4	4	0	4	2	0	3	4	0	9	8	6	5	28
FR(SC) AS=1	21	21	21	29	29	21	21	29	20	29	29	20					
FR(SC) AS=2	19	19	11	22	19	15	19	19	0	19	19	14					
FR(SC) AS=3	17	11	11	17	17	13	17	11	0	17	17	11					
FR(SC) AS=4	14	0	9	14	14	11	14	9	0	9	14	0					
FR(SC) AS=WAS	284	185	200	310	297	240	284	241	56	269	297	155					

In the next step, we calculated rank order correlations between the FRSs per SC with the means from the typicality meta-analysis and with the typicality data from De Deyne et al. (2008). The correlations are in .

Table 72 and the regression plots in Figure 88. The correlations are between .29 and .67 for *fruit,* approximately two thirds of them are > .5. For *vegetable*, they are between -.11 to .69 *with only 25% > .5*. The correlations are higher for both categories when properties from the SC level are used, where for *fruit* the highest is .67 compared to .51 with properties on the C level and for *vegetable* .69 compared to .4 on the C level. Between the different ASs there is no consistent difference: for fruit, the WAS, AS1 the AS3 have the highest results, and for vegetables the AS4 has the best results (.61 to .69) and AS2, AS3 and WAS have comparable results with correlations between .48 and .52. Which typicality dataset is used only makes a small differences except for *vegetable* with C properties, where the mean difference between the correlations with De Deyne et al. (2008) typicality data and the correlation with the mean from the typicality meta-analysis is .11.

This analysis confirms that SC properties have a higher success in predicting typicality than C properties. The observations that participants disagree on applicability in most cases and that the different applicability thresholds do not lead to consistently higher or lower correlations and instead different thresholds produce the best results for the different categories sheds doubt on the utility of applicability for typicality prediction.

Table 72: Correlations between calculated FRS with 5 criteria for applicability (AS1-4 and WAS), two sources of properties (SC, C) and mean from meta-analysis (meta) and mean from De Deyne et al. (2008) typicality (typ) data for fruit and vegetables.

category	properties	typicality data	AS1	AS2	AS3	AS4	WAS
fruit	SC	De Deyne et al.	0.52	0.58	0.58	0.41	0.67
		mean meta-analysis	0.52	0.55	0.58	0.38	0.66
	C	De Deyne et al.	0.43	0.29	0.38	0.49	0.47
		mean meta-analysis	0.45	0.33	0.35	0.51	0.47
		De Deyne et al. vs. meta	-0.02				
		SC vs C	0.08				
vegetables	SC	De Deyne et al.	0.23	0.48	0.50	0.61	0.48
		mean meta-analysis	0.21	0.46	0.52	0.69	0.49
	C	De Deyne et al.	0.40	0.18	0.11	-0.08	0.13
		mean meta-analysis	0.30	0.04	-0.04	-0.11	0.02
		De Deyne et al. vs. meta	0.06				
		SC vs C	-0.13				

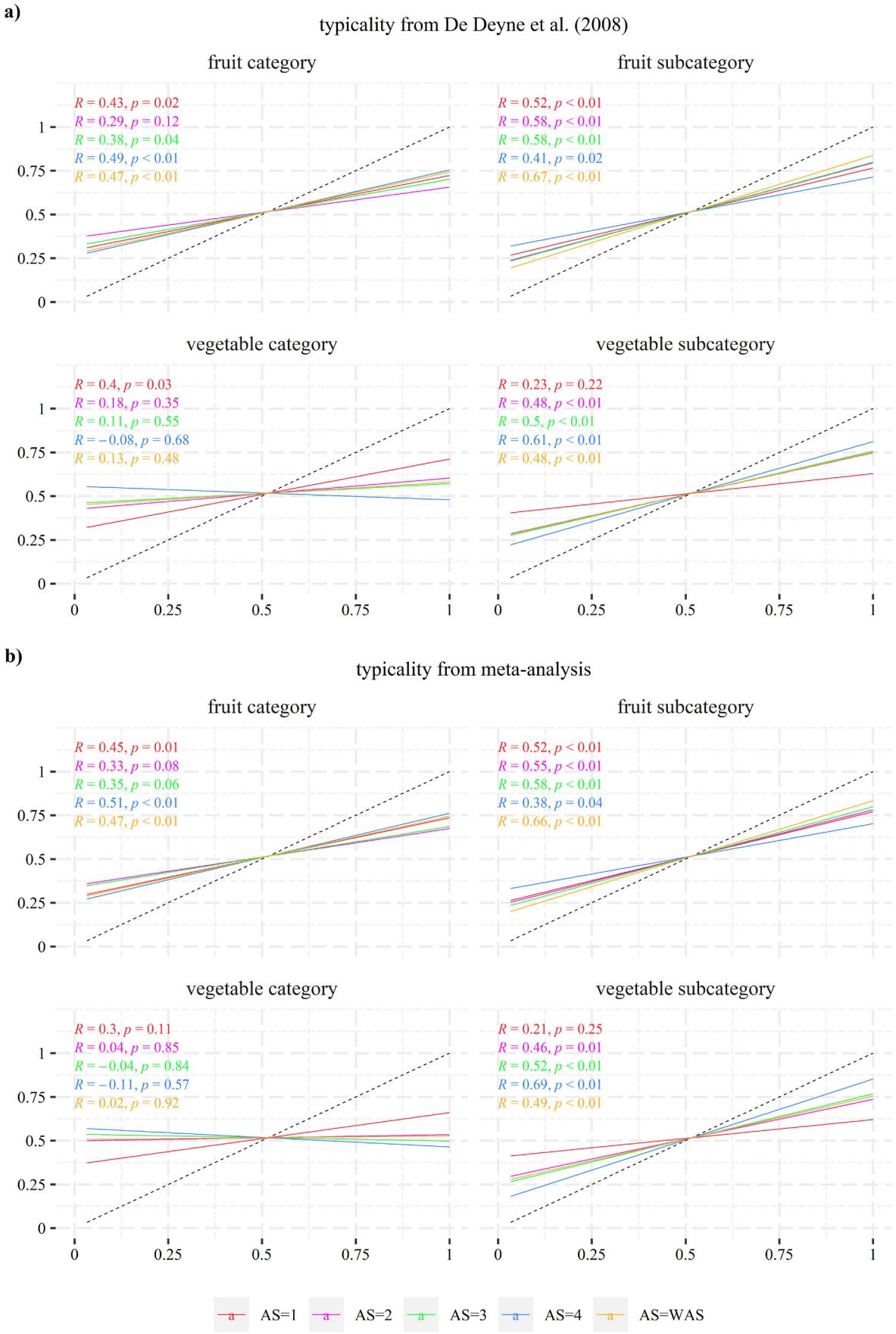

Figure 88: Linear regression plots for the typicality predictions with the FRS with De Deyne et al. (2008) data with typicality a) from the same dataset, b) mean from meta-analysis.

8.1.2 **Probability data**

It is also interesting to see how well probability can be used as an applicability criterion. The probability that an SC has a certain property, gathered in experiment 3, can be interpreted as the applicability of this property for the SC: if the SC has the property with a high probability, the property is applicable to that SC and if the SC has the property with a low probability, the property is not applicable to the SC. An AS can thus be derived from the probability ratings that were collected in experiment 3. FRSs were calculated for three applicability thresholds: the first one is that the mean of the probability ratings of the property for the SC is low, but greater than 0 (>0.1), the second that it is greater than chance probability (>0.5) and the third that it is high (>0.8). These three thresholds are based on different interpretations of applicability: is a property already applicable if it occurs only in some or does it have to occur in at least half or even most of the objects?

FRSs were calculated for both categories with all data and cleaned data and with category properties or all properties and then correlated with the mean typicality from experiment 3.

Table 73 shows example calculations for *fruit*. The FRS per property is the count of all SCs that have a mean probability greater than the respective threshold. The FRS of the SC is the sum of all counts per property multiplied with the properties' FRSs.

The Spearman correlations for all 3 applicability criteria with the mean typicality ratings from experiment 3 are in Table 74 and the regression plots are in Figure 89. The correlations for *fruit* are between -.05 and .85. The highest correlations are found with the high AS criterion, irrespective of whether cleaned or all data were used[36]. For medium and low AS criteria, the correlations increase substantially with cleaning (.42 resp. .49 with all data vs. .69 resp. .65 with cleaned data for the medium level and .14 resp. -.05 with all data vs. .63 resp. .56 with cleaned data for the low level). The correlations for *vegetable* are between -.16 and .53. Here, the highest correlation (.53) is found with the medium AS criterion, category properties and with cleaned data. Cleaning also has a marked effect on the correlations for medium and low AS criteria (.29 resp. .19 with all data vs. .53 resp. .43 with cleaned data for the medium level and -.13 resp. -.16 vs. .20 resp. .17 for the low level).

[36] Most of the properties that were excluded in cleaning had a mean probability <80 and were therefore not included in the calculations for the high applicability criterion.

Table 73: SC-property-pairs and their probability thresholds for fruit.

subcategory	will be made into juice	will be eaten as dessert	grew on a tree	grew on a plant	is juicy	has sugar	has pips/seeds	is sweet	tastes tart	FR(SC) AS low	FR(SC) AS medium	FR(SC) AS high
strawberry	>0.5	>0.8		>0.8	>0.8	>0.8	>0.5	>0.8	>0.1	176	154	110
apple	>0.8	>0.5	>0.8	>0.1	>0.5	>0.8	>0.8	>0.8	>0.1	194	150	106
mango	>0.5	>0.8	>0.8	>0.5	>0.8	>0.8	>0.5	>0.8	>0.1	194	172	106
grape	>0.5	>0.5	>0.1	>0.8	>0.8	>0.8	>0.5	>0.8	>0.1	194	154	88
pineapple	>0.8	>0.5	>0.5	>0.5	>0.8	>0.8	>0.1	>0.8	>0.1	194	150	88
watermelon	>0.5	>0.5	>0.1	>0.5	>0.8	>0.8	>0.8	>0.8	>0.1	194	154	88
peach	>0.5	>0.5	>0.8	>0.5	>0.8	>0.8	>0.5	>0.8	>0.1	194	172	84
tomato	>0.5	>0.1		>0.8	>0.8	>0.5	>0.8	>0.1	>0.1	176	110	66
lime	>0.5	>0.1	>0.8	>0.1	>0.8	>0.5	>0.5	>0.1	>0.8	194	128	62
blackberry	>0.5	>0.5	>0.1	>0.8	>0.5	>0.8	>0.5	>0.5	>0.5	194	176	44
plum	>0.5	>0.5	>0.8	>0.1	>0.5	>0.8	>0.5	>0.5	>0.1	194	150	40
blueberry	>0.5	>0.5	>0.1	>0.8	>0.5	>0.5	>0.5	>0.5	>0.5	194	176	22
passion fruit	>0.5	>0.5	>0.5	>0.5	>0.5	>0.5	>0.8	>0.5	>0.5	194	194	22
pomegranate	>0.5	>0.5	>0.5	>0.5	>0.5	>0.5	>0.8	>0.5	>0.1	194	172	22
pumpkin	>0.1	>0.5		>0.5	>0.1	>0.5	>0.8	>0.1	>0.1	176	88	22
banana	>0.1	>0.5	>0.5	>0.1	>0.1	>0.8	>0.1	>0.5	>0.1	194	84	22
coconut	>0.5	>0.5	>0.8	>0.1	>0.5	>0.5	>0.1	>0.5	>0.1	194	128	18
avocado	>0.1	>0.1	>0.5	>0.5	>0.1	>0.1	>0.5	>0.1	>0.1	194	62	0
fig	>0.1	>0.5	>0.5	>0.5	>0.1	>0.5	>0.5	>0.5	>0.1	194	128	0
papaya	>0.5	>0.5	>0.5	>0.5	>0.5	>0.5	>0.5	>0.5	>0.1	194	172	0
prune	>0.5	>0.5	>0.5	>0.5	>0.1	>0.5	>0.5	>0.5	>0.1	194	150	0
rhubarb	>0.1	>0.5		>0.5	>0.1	>0.5	>0.1	>0.1	>0.5	176	88	0
FR(P) AS low	22	22	18	22	22	22	22	22	22			
FR(P) AS medium	17	19	14	17	16	21	18	17	5			
FR(P) AS high	2	2	6	5	8	10	6	7	1			

Table 74: Spearman correlations of typicality from experiment 3 with typicality predicted from probability data with 3 different interpretations of applicability (high, medium, low).

category	data	properties	high	medium	low
fruit	all	category	.85	.42	.14
		all	.82	.49	-.05
	cleaned	category	.85	.69	.63
		all	.82	.65	.56
vegetables	all	category	.35	.29	-.13
		all	.29	.19	-.16
	cleaned	category	.35	.53	.20
		all	.29	.43	.17

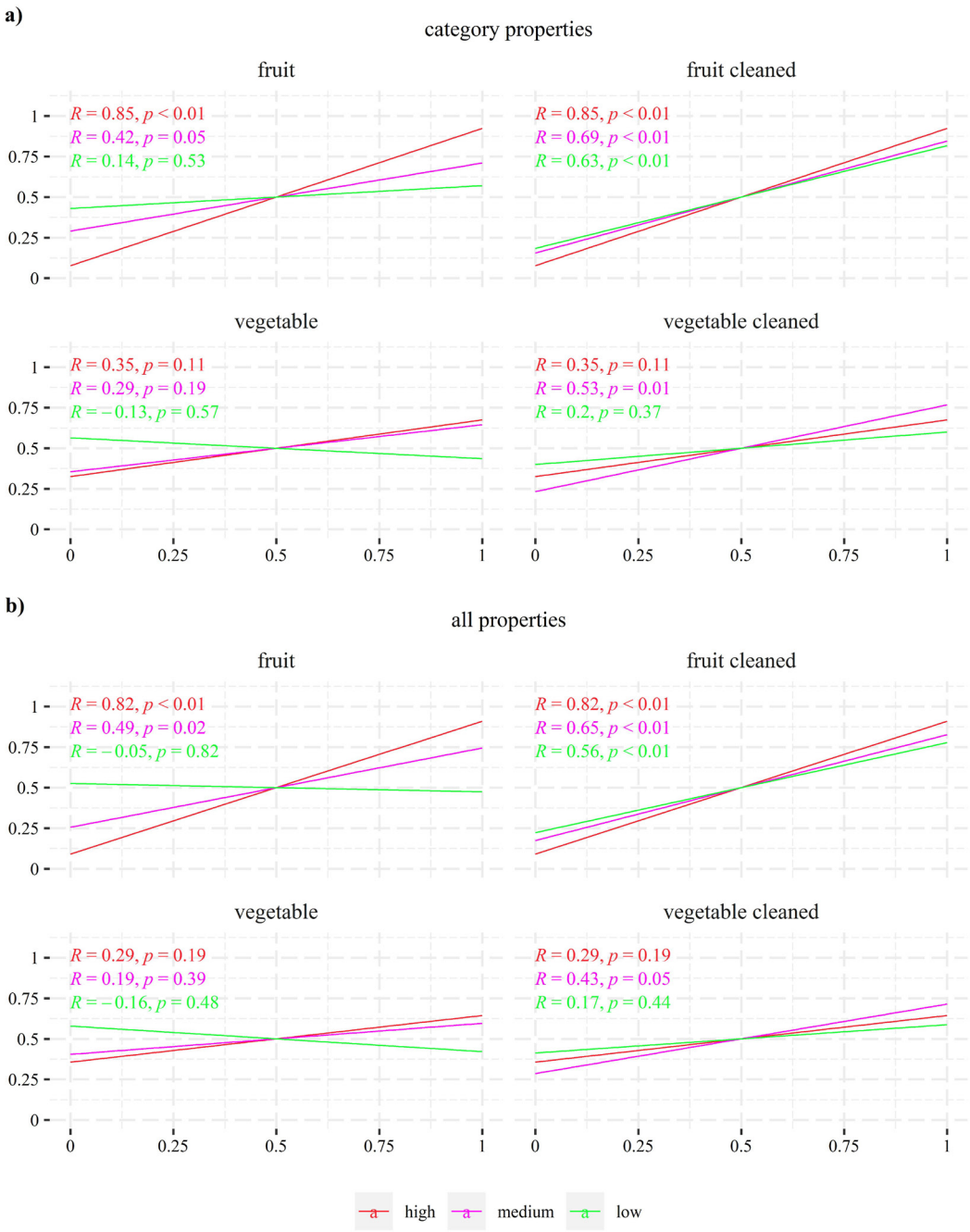

Figure 89: Regression plots for FRS predictions with probability data for fruit and vegetables for 3 applicability thresholds (high, medium, low) with a) category properties only, b) all properties.

Again, the applicability threshold with the best results is not constant for the different categories, which shows once more that applicability is an instable variable. The results for fruit with category properties and a high applicability threshold have correlations in the same order of magnitude as the original experiment reports. Probability ratings can, thus, replace applicability ratings in the right circumstances. The fact that the use of cleaned data increased the correlations consistently can be taken as confirmation that a suitable cleaning procedure for this data type has been found.

8.2 Contrast model

Smith et al. propose to predict typicality (interpreted as similarity Sim between prototype P and instance I) with a version of Tversky's (1977) contrast model (p. 500):

$$\mathrm{Sim}(P,I) = \sum_i v_i \sum_j \left[a \min\left(n_{ij}(P), n_{ij}(I)\right) \dot{-} b\left(n_{ij}(P) - n_{ij}(I)\right) \dot{-} c\left(n_{ij}(I) - n_{ij}(P)\right)\right],$$

where v_i is the diagnosticity of attribute i, a, b and c are parameters and n_{ij} is the number of votes for value j of attribute i. The sign $\dot{-}$ signifies that only positive values are subtracted, and negative values are set to 0. The diagnosticity weights for the attributes were calculated as the v-statistics:

$$\mathrm{diag}(A_i) = v = \sqrt{\frac{\chi^2_{A_i}}{N_{A_i}}},$$

where $\chi^2_{A_i}$ is the chi-squared statistics for all the values named per attribute and N_{A_i} is the total number of votes for the attribute.

The parameters a, b and c were fitted with a software to obtain maximal correlations with typicality data. They note that "the parameters are similar for the two kinds of concepts, and the ordering of the parameters is in agreement with prior results (Gati & Tversky, 1984, Tversky, 1977; Tversky & Gati, 1982, Tversky & Gati, 1982)" (p.504) and

> "The three parameters of the contrast model are very stable. In virtually every case, the weight given to common features, a, exceeds that given to features distinct to the concept [i.e., the prototype], b, which in turn exceeds that given to features distinct to the instance [i.e., the SC], c." (pp. 515-516).

Smith et al. asked 30 participants to name properties for 15 vegetable or fruit instances, deleted all properties that were only mentioned once and then let 2 judges intuitively decide which properties are values of the same attribute and deleted all about which there was disagreement. They report a correlation of .75 for fruit and .4 for vegetable between typicality predicted with their formula and rated typicality. They explain the low value for vegetable with the fact that many SCs they included have a typicality on the medium or high typicality level which leads to small differences between the SCs.

The application of their formula to new datasets requires information on the PF of attribute values which are available from the McRae et al. data set used as a basis for the stimuli of experiment 1 (section 6.1.2). The properties they collected have to be decomposed into attributes and their values to apply the formula. I present an application of their formula to the McRae dataset in section 8.2.1 and to the probability data gathered in my experiments in section 8.2.2.

8.2.1 McRae et al. data

The structure of the dataset from McRae et al. (2005) is described in detail in section 6.1.2. The procedure proposed by Smith et al. can be applied to it well, because it contains productive frequency (PF) data for each property. All properties that were generated for each SC for which "a fruit" or "a vegetable" was mentioned were filtered from the original dataset. Then, I did an attribute-value assignment to each property, relying closely on the attributes that were used by

Smith et al. Some properties[37] were excluded because they did not fit into their attribute-value structure. Some properties had to be merged because they had identical attribute-value-assignments[38]. Then, attribute diagnosticity was calculated as v-statistics. Table 75 shows the calculation of the v-statistics for the OUTSIDE-COLOUR attribute as an example. The v-statistics is the square root of χ^2 divided by the total number of observations in the table. It reflects the difference between observed frequencies in the cells and the expected frequencies that would be observed in an equal distribution between groups. The expected value E_{ij} for the cell in row i and column j is defined as the row total R_i multiplied with the column total C_j divided by the grand total N:

$$E_{ij} = \frac{R_i C_j}{N}.$$

For example, the expected value of the colour black for *fruit* is

$$E_{Black,fruit} = \frac{62 \cdot 772}{1{,}462} = 32.7,$$

a bit lower than the observed productive frequency 36, which indicates that the value is slightly discriminative for the two categories. The squared differences between observed and expected frequencies are divided by the expected frequency and then summed up to calculate χ^2. The diagnosticity for SMELL and WHEN-GROWN was set to 1, because they only have one row with votes greater than 0 in only one column (SMELL-Strong has only votes for *vegetable* and WHEN-GROWN-In Summer has only votes for *fruit*), which leads to a division by 0, thus the χ^2-statistics is not computable. The decision to set them to 1 was made to reflect Smith et al.'s thought that

> "the diagnosticity of an attribute would be largely a matter of how useful it was for discriminating between fruits and vegetables. […] This statistic varies between 0 and 1 and indicates the extent to which the values of the attribute are associated with *fruit* but not *vegetable*, or vice versa." (p. 498)

The fact that the attributes have only votes for one category makes them very discriminative for that category.

[37] "eaten in summer", "used in autumn" "eaten at Christmas/Thanksgiving", "eaten for breakfast", "like a lemon/peach/potato/cucumber", "is edible/harvested/picked", "is tropical", "comes frozen", "associated with Popeye" and everything referring to super- and subcategories.

[38] "has green skin" and "is green" to OUTSIDE-COLOUR-Green for cucumber, "is green" and "has a green outside to OUTSIDE-COLOUR-Green for zucchini, "is white" and "is white inside" to INSIDE-COLOUR-White for coconut, "has lots of water inside" and "has water" to JUICINESS-Watery for lettuce, "made from grapes" and "made from dried grapes" to ORIGINAL-IDENTITY-Grapes for raisin, "is brown" and "has a brown outside" to OUTSIDE-COLOUR-Brown for coconut, "is red" and "has a red outside" to OUTSIDE-COLOUR-Red for radish.

Table 75: Example calculation of diagnosticity as v for the OUTSIDE-COLOUR attribute from the McRae et al. dataset.

Attribute	Value	PF fruit (O)	PF veg (O)	total	Expected fruit (E)	Expected vegetable (E)	$(O-E)^2/E$ fruit	$(O-E)^2/E$ vegetable
	Black	36	26	62	32.7	29.3	0.3	0.4
	Blue	27	0	27	14.3	12.7	11.4	12.7
	Bright	5	0	5	2.6	2.4	2.1	2.4
	Brown	31	30	61	32.2	28.8	0.0	0.1
	Dark	5	0	5	2.6	2.4	2.1	2.4
OUTSIDE-COLOUR	Green	180	405	585	308.9	276.1	53.8	60.2
	Orange	117	66	183	96.6	86.4	4.3	4.8
	Pink	13	0	13	6.9	6.1	5.5	6.1
	Purple	46	30	76	40.1	35.9	0.9	1.0
	Red	188	91	279	147.3	131.7	11.2	12.6
	Yellow	124	42	166	87.7	78.3	15.1	16.9
	total	772	690	1462				
						$\chi^2 = \sum \frac{(O-E)^2}{E}$	226.1	
						$v = \sqrt{\frac{\chi^2}{N}}$	0.39	

Table 76 compares the results of my calculations and assignments with the results reported in Smith et al. The total number of votes was much higher for the McRae et al. dataset[39], because 29 fruit and 27 vegetable SCs were available compared to 15 each in the original paper. Therefore, the differences in the number of votes were calculated in terms of the percentages of all votes. The differences are all under 10%. However, the differences in diagnosticity are high: >|0.2| for 50% of the attributes. This means that while the relative votes per attribute between the two datasets are similar, the votes per value differ which can be seen in the different results of the χ^2 statistic. This is either due to me assigning values in a different way or due to the addition of SCs that could have changed the distribution of votes over values, but not the number of votes per attribute. As they did not publish their value assignments, the reason for this difference cannot be determined. Considering that their original sample contained only 15 SCs it is not unreasonable to assume that the attribute-value-structure they generated was not representative of the category as a whole.

[39] There are 2,606 votes for *fruit* and 2,192 votes for *vegetable* in the original article and 3,880 votes for *fruit* and 3,186 votes for *vegetable* in the McRae et al. data.

Table 76: Comparison of data in Smith et al. (1988, p. 500) with data derived from McRae et al. (2005): total votes for fruit and vegetables with differences in% of total votes and their diagnosticity with difference.

	Smith et al. Data			McRae et al. Data					
Attribute	Total Votes for Fruit	Total Votes for Vegetables	Diagnosticity v	Total Votes for Fruit	Diff in % of total votes	Total Votes for Vegetables	Diff in % of total votes	Diagnosticity v	Diff
OUTSIDE-COLOUR	503	462	0.44	772	-0.6	690	-0.6	0.39	-0.05
OUTSIDE-TEXTURE	261	206	0.61	92	7.6	124	5.5	0.91	0.30
TASTE	252	167	0.68	488	-2.9	147	3.0	0.62	-0.06
HOW-EATEN	238	397	0.84	480	-3.2	626	-1.5	0.80	-0.04
HOW-GROWN	203	90	0.82	306	-0.1	149	-0.6	0.84	0.02
SEEDS	191	48	0.13	204	2.1	103	-1.0	0.14	0.01
SHAPE	157	158	0.70	230	0.1	170	1.9	0.41	-0.29
JUICINESS	146	34	0.61	240	-0.6	38	0.4	0.65	0.04
INSIDE-COLOUR	119	30	0.37	45	3.4	92	-1.5	0.59	0.22
SIZE	109	34	0.12	170	-0.2	68	-0.6	0.10	-0.02
PIT	55	0	0.43	123	-1.1	18	-0.6	0.27	-0.16
INSIDE-TEXTURE	51	44	0.52	126	-1.3	161	-3.0	0.70	0.18
ORIGINAL IDENTITY	44	18	0.55	44	0.6	19	0.2	1.00	0.45
WHERE-GROWN	41	19	0.96	107	-1.2	173	-4.6	0.76	-0.20
SKIN	39	25	0.50	204	-3.8	66	-0.9	0.72	0.22
STEM	38	15	0.24	23	0.9	33	-0.4	0.83	0.59
VARIETIES	37	83	0.71	10	1.2	34	2.7	1.00	0.29
SIDE-EFFECTS	27	47	0.95	29	0.3	88	-0.6	0.92	-0.03
WHEN-GROWN	22	2	0.34	5	0.7	0	0.1	1.00	0.66
CONTAINER	21	12	0.53	16	0.4	14	0.1	1.00	0.47
NUTRITIONAL-VALUE	18	138	0.83	64	-1.0	147	1.7	0.68	-0.15
LEAF	14	68	0.67	29	-0.2	98	0.0	0.33	-0.34
FAVOURITE-CONSUMER	9	47	0.90	18	-0.1	19	1.5	1.00	0.10
NONFOOD-USE	7	10	0.79	50	-1.0	79	-2.0	0.57	-0.22
COST	4	2	1.00	5	0.0	5	-0.1	1.00	0.00
SMELL	0	36	0.41	0	0.0	25	0.9	1.00	0.59

The attribute-value-structure for the fruit and vegetable category (or prototype) was determined by "averaging over all relevant instances on each attribute" (p.498), meaning by taking the mean number of votes per category for each attribute.

To estimate the parameters a, b and c, Smith et al. used the software STEPIT (Chandler, 1961) with the goal to maximise the correlations between rated and predicted typicality. I computed the correlations with the mean from the typicality meta-analysis and with all datasets that had an overlap of at least 20 SCs with the McRae et al. data. The parameters were fitted to maximise the sum of all correlations with the Excel Solver, setting the parameters to starting values reported in their article: a = 1.84, b = .5 and c = .2 for *fruit* and a = .88, b = .5 and c = .2 for *vegetable* (p. 515). Because of the high difference in diagnosticity, the results were also computed using the ones reported in Smith et al. (1988, p. 500).

The mean correlation for *fruit* (Table 77) is high (.77) and increases only slightly (+.02) with the fitted parameters a = 0.56, b = 1.24 and c = 0.05. For *vegetable* (Table 78), the mean

correlation is considerably higher with the fitted parameters a = 2.11, b = 0 and c = 0.06: the mean rises from .34 (SD = .24) to .59 (SD = .17). Notably, the correlation with the typicality means reported in Smith et al. is the highest both for fruit (.9) and for vegetable (.77).

Table 77: Parameters and correlations between typicality from 7 datasets and predictions from contrast model based on PF-data from McRae et al. 2005 for fruit.

		a=1.84, b=0.5,c=0.2		a=0.56, b=1.24, c=0.05	
dataset	n	r	p	r	p
Meta mean	29	.81	.000	.82	.000
Rosch 1975	27	.81	.000	.79	.000
Uyeda et al. 1980	23	.78	.000	.80	.000
Hampton et al. 1983	25	.75	.000	.76	.000
Brown et al. 1986	26	.71	.000	.76	.000
Schwanenflugel et al. 1986 ENG	20	.67	.001	.74	.000
Schwanenflugel et al. 1986 SPA	20	.75	.000	.79	.000
Smith et al. 1988	11	.90	.000	.91	.000
mean		.77		.79	
SD		.07		.05	

Table 78: Parameters and correlations between typicality from 7 datasets and predictions from contrast model based on PF-data from McRae et al. 2005 for vegetables.

		a=0.88, b=0.5,c=0.2		a=2.11, b=0, c=0.06	
dataset	n	r	p	r	p
Meta mean	28	.40	.035	.65	.000
Rosch 1975	25	.40	.045	.56	.003
Uyeda et al. 1980	20	.00	.998	.30	.205
Hampton et al. 1983	24	.24	.269	.64	.001
Brown et al. 1986	24	.14	.527	.50	.012
Smith et al. 1988	10	.54	.106	.77	.010
Ruts et al. 2004	20	.62	.004	.75	.000
De Deyne et al. 2008 TYP	20	.66	.002	.72	.000
De Deyne et al. 2008 GDN	20	.59	.006	.73	.000
Schröder et al. 2012	20	.19	.430	.55	.011
Moreno-Martinez et al. 2014	20	.02	.942	.28	.230
mean		.34		.59	
SD		.24		.17	

As the diagnosticity values calculated based on the McRae et al. data were differing from the ones reported in Smith et al., all correlations were also calculated with the values they reported. This made no considerable difference in the correlations (<|0.05|) for *fruit*. For *vegetable*, the correlations increased slightly, in two cases +.06, and one case each +.07 and +.08.

In this analysis, I was able to reproduce the original results from Smith et al. The correlations for both categories were even higher for most models, even with the use of the original parameters that were fitted on a different dataset. The fitted parameters deviate from the ones reported in Smith et al. which contradicts their observation that they are stable and similar for both categories. While fitting the parameters only made a minor difference for fruit, it raises

the correlations for vegetable considerably. The dependence of the model on unstable parameters leads to it taking an extra step for typicality prediction.

8.2.2 **Probability data**

I also used the probability data from my experiments as an input for the contrast model. I used the probabilities gathered in experiment 2 and 3 to replace the number of votes. Two versions of a category prototype were computed: in the first, the mean probability estimations for the fruit and vegetable category from experiment 2 were used as an input, in the second, the average probability of all SCs from experiment 3 was used, analogous to the procedure in Smith et al. Figure 90 compares the two. The mean estimated category probability has higher or equal values as the mean SC probability.

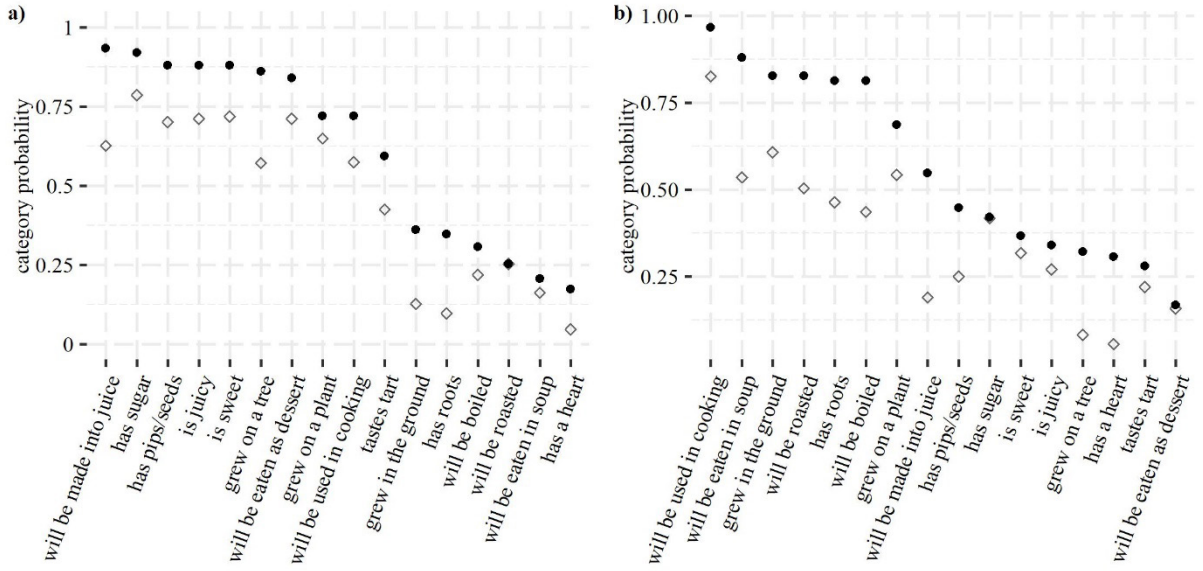

Figure 90: Category probability from experiment 2 (dots) and average category probability from experiment 3(diamonds) for a) fruit and b) vegetables.

The contrast model was computed with the diagnosticities reported in Smith et al. and with the diagnosticities[40] from the models in 7.1 with category properties (termed diag sep) and all properties (termed diag uni). As the attributes used were not identical (see section 7.1.1), they had to be assigned equivalents to be comparable: I used the diagnosticity of their attribute HOW-EATEN for CONSUMED-IN, PREPARATION-METHOD and USED-IN-COOKING. Table 79 and Figure 91 compare the three available diagnosticities. For *fruit*, the highest difference is in the diagnosticity of PREPARATION-METHOD, which has a very high diagnosticity in Smith et al. (.84) and a low one in our data (.24). Similarly big differences are found for ROOTS and SEEDS. For vegetable, the highest difference is for MAIN-NUTRITIONAL-COMPONENT, which has a high diagnosticity in Smith et al. (.83) and a low one in our data (.26). It should however be noted that Smith et al. used the same diagnosticity values for both categories, while I calculated them separately for each category. PREPARATION-METHOD has a high diagnosticity for *vegetables* in our data and MAIN-NUTRITIONAL-COMPONENT has a high diagnosticity for *fruit*, which means a low diagnosticity in the respective contrast category.

[40] I did not normalise the diagnosticities and used the maximum value per attribute instead, because the results of the contrast model are not normalised.

Table 79: Attributes from experiment 3 and from Smith et al. with their diagnosticities for fruit and vegetables.

Attribute Exp 3	Attribute Smith et al.	Smith et al. diag	diag fruit sep	diag fruit uni	diag veg sep	diag veg uni
CONSUMED-IN	HOW-EATEN	.84	.79	.90	-	.87
TASTE	TASTE	.68	.54	.77	-	.25
JUICINESS	JUICINESS	.61	.54	.77	-	.23
MAIN-NUTRITIONAL-COMPONENT	NUTRITIONAL-VALUE	.83	.47	.74	-	.26
HOW-GROWN	HOW-GROWN	.82	.45	.72	.55	.77
SEEDS	SEEDS	.13	.43	.71	-	.29
HEART	INSIDE-TEXTURE	.52	-	.37	.27	.63
PREPARATION-METHOD	HOW-EATEN	.84	-	.24	.63	.82
ROOTS	OUTSIDE-TEXTURE	.61	-	.22	.56	.78
USED-IN-COOKING	HOW-EATEN	.84	-	.40	.21	.60

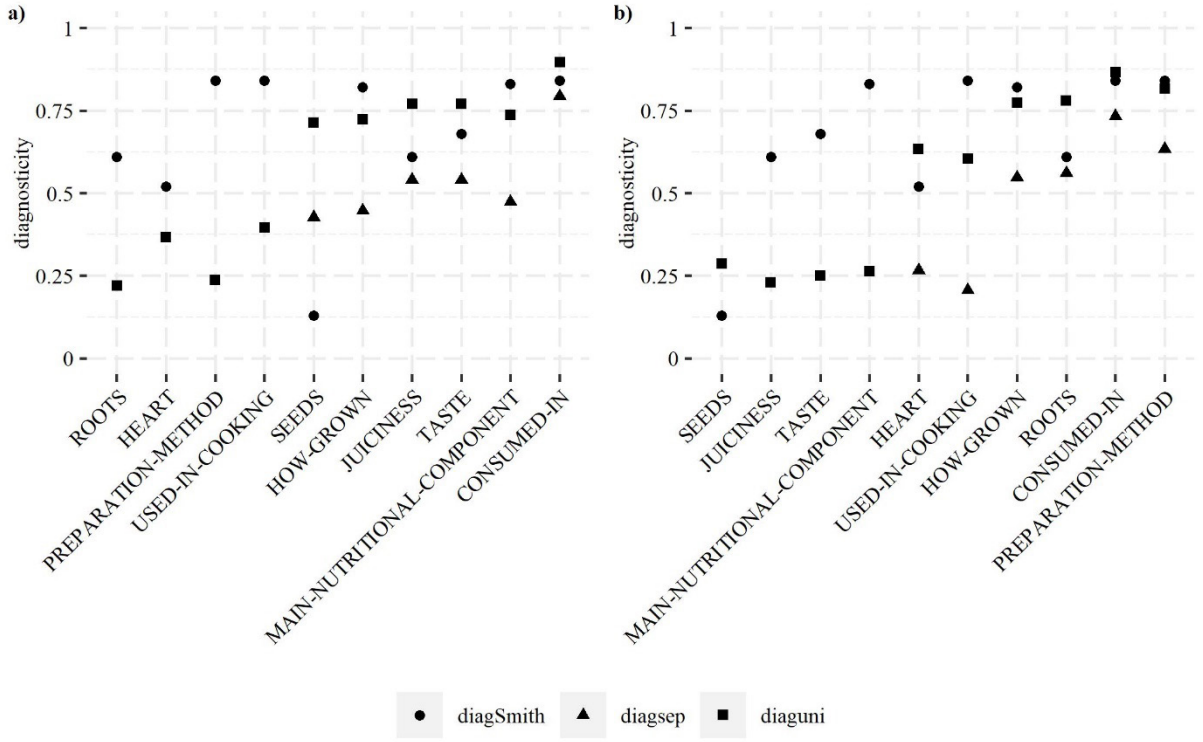

Figure 91: Diagnosticity per attribute from Smith et al. (dots), probabilistic prototypes with category properties (diag sep, triangles) and with all properties (diag uni, squares) for a) fruit and b) vegetable.

The correlations with Smith et al. parameters as well as fitted parameters for all three diagnosticity values are in

Table 80. Again, the predictions were computed with all data as well as with cleaned data. For *fruit*, all correlations are > .5. The lowest correlations are found for predictions with the diagnosticity weights from Smith et al. (v). Parameter-fitting (r max) only increases them insignificantly (+.03 to +.04), but the use of cleaned data raises them (+.12) to .66. The predictions with the diagnosticity from experiment 1 have higher correlations with typicality. With Smith et al. parameters, the predictions with uncleaned data have higher correlations (.84

resp. .71) than those with cleaned data (.78 resp. .71). With fitted parameters, the predictions with cleaned data have only slightly higher correlations (.89 resp. .84) than the predictions with all data (.84 resp. .82). The inclusion of contrast category properties does not make a substantial difference for the correlations, but for the tendency of the fitted parameters. For category properties with all data, parameter fitting did not change the correlation substantially. For category properties with cleaned data, setting the parameters to a = .03, b = 1.2, c = 0 raised the correlation by +.11 to .89. In this constellation, the highest weight is given to the excess probability of the prototype. For predictions with all properties, the highest weight is on c, the excess probabilities of the SCs, both in the cleaned and uncleaned case. For *vegetable*, all correlations are >.4. The predictions with cleaned data are higher correlated with typicality than those with uncleaned data (.61-.64 vs. .43-.56). Here, the predictions with all properties have much higher correlations after fitting such that the excess probability of the SCs has the highest weight. The highest correlation (.72) is found for all properties, cleaned data and parameters fitted to a = .2, b = .1, c = .4.

The correlations are in the same order of magnitude as and in some cases even higher than the ones reported in the original study. Subjective probabilities are thus a suitable replacement for the number of votes as a measure of typicality contribution of properties in the contrast model. As noted in the previous section, the fitted parameters vary largely between the different conditions and therefore do not seem to be universal as assumed in the original paper.

Table 80: Correlations, p-values between predictions of the contrast model with probability data from experiment 1, 2 and 3 and various typicality data for a) fruit and b) vegetable.

a) fruit

				fitted parameters		
diag	cleaning	r	r max	a	b	c
v	all	.54	.58	.6	.4	.4
	cleaned	.66	.69	.7	.4	.3
category properties	all	.84	.84	2.0	.6	.0
	cleaned	.78	.89	.3	1.2	.0
all properties	all	.76	.82	.1	.4	.5
	cleaned	.71	.84	.2	.3	.5

b) vegetable

				fitted parameters		
diag	cleaning	r	r max	a	b	c
v	all	.42	.48	.4	.3	.4
	cleaned	.61	.67	.5	.3	.5
category properties	all	.54	.54	.2	.3	.4
	cleaned	.62	.66	.1	.0	.6
all properties	all	.48	.70	.0	.2	.5
	cleaned	.59	.72	.1	.1	.5

8.3 **Discussion**

Table 81 summarises the highest correlations of mean SC typicalities with predictions per model and the data used as input. Probabilistic prototypes could only be used with the data gathered in experiment 1 to 3, as I am the first to collect subjective probability ratings for properties in the required systematic manner. The FRS could be replicated with the De Deyne et al. data that contain application scores and with the data from experiment 1-3, when

thresholds for mean probability ratings are defined as applicability. The predictions for *fruit* with experiment 1-3 data predict typicality in the same order of magnitude as the originally reported correlations, while the data for *vegetable* as well as the predictions with the De Deyne et al. application scores are clearly below them. The results from the contrast model in Smith et al.'s version could be replicated with the data from McRae et al., which contain productive frequency data, as well as with the data from experiment 1-3, when the number of votes is replaced with mean subjective probability. The replications had higher correlations with typicality than the original data, which, as was already noted in their article, did not include enough different SCs.

Table 81: Summary of the highest predictions per model and used data.

model	data	fruit	vegetables
Probabilistic prototypes	experiment 1-3	.84	.68
FR score	Rosch and Mervis	.85	.84
	De Deyne et al.	.67	.69
	experiment 1-3	.85	.53
Contrast model	Smith et al.	.75	.40
	McRae et al.	.79	.75
	experiment 1-3	.89	.72

The components of all models are compared in Table 82. While the FRS uses the simple property list format to represent concepts, the contrast model and probabilistic prototypes use the more fine-grained format frames. Both the FRS and the contrast model incorporate randomly determined numbers – the maximally possible FRS depends on the number of SCs and properties that are used in the calculation and the maximally possible typicality predicted by the contrast model depends on the number of properties and the number of participants, as they determine the maximally possible PF per property. As both numbers are not a priori restricted, both of their predictions cannot be normalised. Probabilistic prototypes have normalised predictions that lie between 0 and 1, the range of predictions is not dependent on the number of SCs or properties. The number of properties required to make a prediction is much lower for probabilistic prototypes, because only those that are frequent and diagnostic are needed, which was shown to apply to only a fraction of the properties generated for the categories. If data on additional properties is of interest, the probabilistic prototype can easily be extended and reflect their typicality contribution.

Table 82: Comparison of the components of FR score, contrast model and probabilistic prototypes.

	FR score	Contrast model	Probabilistic prototypes
representation format	property list	frame	frame
property contribution	applicability $AS(SC_i) \in \{0,1\}$	productive frequency $PF \in [0, N]$ $N \mathrel{\widehat{=}}$ number of participants	prototypicality i.w.s. $Pr(V\|C) \in [0,1]$
weight	properties' FR score $FR_{P_k} \in [0, N]$ $N \mathrel{\widehat{=}}$ number of SCs	diagnosticity $diag(A_i) \in [0,1]$	prototypicality i.n.s. $diag(A_i) \in [0,1]$
formula	weighted sum	weighted sum	weighted average
prediction	$typ \in [0, N]$ $N \mathrel{\widehat{=}}$ number of SCs and number of properties	$typ \in [0, N]$ $N \mathrel{\widehat{=}}$ number of properties and number of participants	$typ \in [0,1]$
amount of properties	all produced	all produced	only diagnostic and frequent, but extensible

I have argued in section 4.3 and shown in this chapter that the FRS's dependence on the notion of applicability is problematic. In the data from De Deyne et al., participants did not systematically agree which properties are applicable to which SC and the "degree" of applicability that predicted typicality best varied depending on the category. Similarly, in the predictions with probability data, the highest correlations for *fruit* were found when applicability was interpreted as a high probability that an SC has the property, while for *vegetable* applicability is interpreted as medium probability in the best predictions. The fact that the notion of property applicability is vague and context-dependent renders the results relative to its interpretation.

The use of PF data in the contrast model is problematic as well because PF is presumably biased by the salience of properties, which does not necessarily correspond to the importance in typicality contribution.

It can be argued that, all in all, probabilistic prototype frames are the preferrable alternative because they have a solid theoretical foundation in the laws of probability and evolution, and they rely on data that is easily collected and convertible into the other model's input without loss of predictive power.

9 Conclusion and outlook

This chapter comprises a summary of this thesis (9.1). and an outlook on promising future research directions (9.2).

9.1 Summary

In the foregoing, I presented a theoretical analysis and empirical investigation of parametric models for the prototype theory of concepts.

I began with a demonstration of arguments and phenomena that show that the prototype theory of concepts is essential to account for the meaning and structure of concepts in certain domains in which the classical theory offers unsatisfying accounts.

In a meta-analysis of typicality data, I found that the correlations between typicality orderings from different studies are in general high, throughout different decades, instructions and scale types and, for *fruit*, languages and culture. The *vegetable* typicality orderings are highly correlated within the same language and culture, but low between different ones. Many studies present results that are based on high typicality levels – most SCs for which several mean ratings are available in the literature are located on the high end of the typicality scale. On the medium and low level, I found differences in the interpretation of the meaning of the lowest scale point and many unique SCs. The fact that the studies differ in the interpretation of the lower scale points and in the choice of less typical SCs illustrates an uncertainty about the meaning of low typicality. The SD within and between datasets was highest for SCs with mean ratings on the medium typicality level. Here, multimodal distributions and thus low intersubjective agreement, are prevalent, based on the analysis of four datasets for which all ratings were available. I concluded from this analysis that typicality orderings should not be interpreted as fine-grained as they commonly are, when small mean differences are taken to reflect small typicality differences. Instead, I argued that the data show three clearly different typicality ranks that differ in terms of the number of modes and variance of their rating distributions. It is, however assumed, that the concrete rank orders between all SCs can be determined with a large enough sample size.

Based on my analysis of the constituents of formal models for prototype concepts in chapter 4, I presented a new model for the representation of category concepts and introduced a formula that predicts typicality based on the probabilities and inverse probabilities of a property (or value) given a (sub)category: probabilistic prototype frames that are based on work from Gerhard Schurz (in particular 2012, moreover 2005, 2007, 2011). They incorporate subjective estimations of the probability of properties that contribute to the evolutionary norm state in quantified frames. Then, I presented how I collected the data for their empirical confirmation in three experiments. Even though it is suspected that the participants did not strictly rate probability as frequency, but – depending on the property – as possibility or uncertainty, the ratings were intersubjectively stable for a good amount of the data. Unreasonable SC-property-combinations were either due to the matrix design of the experiment or due to properties about which the participants lacked knowledge, mostly those describing growing conditions; they were recognizable by their high variance, and it was argued that they should be excluded in further use of the data. It was shown that the probability ratings of participants have almost perfect correlations with Bayes' theorem. Furthermore, the mean property probability ratings for SCs from the high and medium typicality level had almost perfect correlations with the mean property probability ratings for Cs.

The data were used to test the ability of probabilistic prototype frames to predict the typicality of subcategories for a category by means of the proposed typicality formula. The correlations were significant and had a high effect size. Using cleaned data, i.e., only those properties that had rating distributions with a low IQR, improved the correlations considerably. Between cleaned models, no significant difference was found between applying diagnosticity weights on the attribute and on the property level or even omitting them. Using all data, the property models had the highest correlations. This observation may however be due to choosing not enough values per attribute which was done to have a cross-categorical design and to test the influence of contrast category properties on the predictions. The inclusion of contrast category properties in the calculations reduced the correlations in all cases and should therefore be avoided in further studies. The lower correlations for *vegetable* were explained by the facts that (1.) the property lists lack typical properties for *lettuce* and *spinach*, which would also be solved by complementing the prototype frame with more values in future experiments, and moreover, that (2.) the prototype for *vegetable* is culturally different in ways that our culturally mixed empirical data might not be able to capture.

I went on to investigate the two most prominent proposals of quantifying typicality: the family resemblance score (FRS) from Rosch and Mervis (1975) and Smith et al. (1988)'s version of the contrast model. I applied their proposals to new datasets and used the probability data gathered in my experiments as their input. The FRS's weakness was shown to lie in the notion of applicability, for which no consistently successful interpretation was found. The contrast model had higher correlations than those reported in the original article, both when applied to the McRae et al. data and to the probability data. For the high result, the parameters had to be fitted into a very different direction for each dataset. This introduces an additional step in the prediction and makes the interpretation of the results more difficult. It is also not consistent with Smith et al.'s observation that the parameters are constant and stable across categories. Furthermore, the use of productive frequency data could introduce a bias due to the large role that salience plays in those.

There is no advantage of probabilistic prototype predictions compared to the best results obtained for family resemblance predictions in terms of correlations between predicted and experimentally obtained mean SC typicalities; the correlations of the contrast model predictions were slightly below ours. Probabilistic prototypes have, however, two main advantages. First, they are theoretically well-founded in probability theory and the generalised theory of evolution which guarantees that the probabilities they incorporate are based on non-accidental properties of the environment. Second, the data they require are straight-forwardly empirically collectable without relying on vague or biased notions. Contrary to the other models, they do not require data on a high number of properties and instead focus on those that were determined to be relevant because they have a high frequency and discriminative power for the category.

9.2 Outlook

The work presented in this thesis is a solid basis for future research in several directions. It provides empirical insights into identifying prototype representations which, in addition to their intrinsic value for semantics, can then be used to explain and predict phenomena like conceptual combination with category concepts (section 9.2.1), as a basis for the embedding of metric information in frames (section 9.2.2), to investigate the intersubjective stability of frames from different individuals for more subjective concepts (section 9.2.3) and to investigate many more concepts for which probabilistic representations are interesting (section 9.2.4). Probabilistic prototypes could be improved by exploring different ways to identify properties (section 9.2.5) and to measure typicality (section 9.2.6).

9.2.1 Modelling conceptual combination and constraints in prototype frames

In Strößner (2020a), a formal, Bayesian model of modification with prototype frames is presented. She complements the Selective Modification Model from Smith et al. (1988) with background knowledge in the form of probabilistic constraints to provide a more complete account of prototype combination that explains why default inheritance can be blocked or weakened by information about the relevance of modifiers (see sections 2.3.3 and 5.3). Figure 92 illustrates her model in theory (top) and for the example of *pet hamster* (bottom). In the Selective Modification Model, modification corresponds to a shift of all the votes of the modified attributes' values to the modified value and raising the diagnosticity of the modified attribute. The first step is the same in her model – the targeted value (V_1 or Pet) of the modified attribute (A_1 or DOMESTICITY) receives maximal probability. In the second step of her model, background beliefs in the shape of probabilistic constraints like $\Pr(W_1|V_1)$ change the probability of relevant attribute values. In her example, the background belief is that pets live with .8 probability in cages and the LIVING-ENVIRONMENT attribute values are updated to reflect this. Her frame for *pet hamster* reflects that pet hamsters live in cages with a higher probability than hamsters, in addition to the obvious property of being a pet. The thought that the importance of properties is separately evaluated in conceptual combinations is also found in Hampton's Composite Prototype Model (Hampton 1987, Hampton and Jönsson 2012).

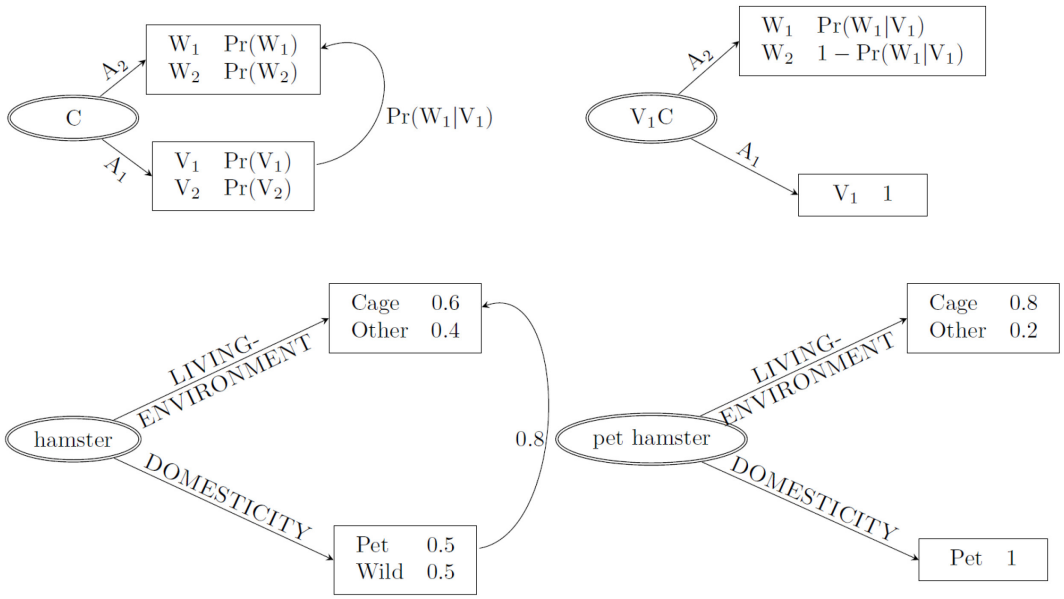

Figure 92: Conceptual modification in theory (top) and for hamster and pet hamster (bottom) following Strößner, 2020a, pp. 868–870.

She shows that this model is easily extensible to recursive frames and that the influence of constraints is global, i.e., they influence all attribute values. Then it is shown how the influence of constraints is constrained in this model: modification with atypical values can have a great influence on the probability distribution, while modification with typical ones cannot.

The predictions of this normative model would be very easily testable with probabilistic prototypes as presented in this thesis. First, the prototype frame for the concept (e.g., *hamster*) would have to be determined in experiments analogous to experiments 1 and 2 of this thesis. Second, the conditional probabilities for all attribute values would have to be collected, for example by asking questions of the type "How probable is it that a hamster has the following properties, if you know that it is wild?" for all possible property combinations. To test whether the constraints affect the modified concept in the predicted way, the prototype frame for the modified concept would then have to be identified and its probabilities compared to the ones predicted by the model.

9.2.2 Representing metric information in frames

The properties in our investigation were all assumed to be discrete. That is, a single object either has this property or not. Many properties, like those referring to size and colour, describe measurable qualities and can therefore take non-discrete or continuous values. While measured quantities can easily be embedded in frames (Schurz & Votsis, 2014, Kornmesser & Schurz, 2020), another interesting possibility is to embed conceptual spaces in frames wherever needed as proposed in Strößner (2020b). She argues that, while the representation of classificatory values is uncommon in conceptual spaces and common in frames, and vice versa for metric values, both are possible in both representation formats. The only two differences between the two that she identifies are that conceptual spaces cannot embed recursivity and that frames cannot be compared with distance measures (p. 692). She takes the criterion C, which is Gärdenfors' proposal to construct concepts by combining a set of regions of domains with weights that reflect information about salience and the correlation between the domains (Gärdenfors, 2000, p. 105), to be what can unite frames and conceptual spaces: "Conceptual spaces can model the inner structure of (natural) properties, but frames allow to link properties to each other and to the concept." (Strößner, 2020b, p. 693). Figure 93 shows an example how this integration could work. The concept *apple* is exemplarily represented with three attributes, two of which, PEEL and FLESH, are recursive. Both attributes are further specified by the attributes COLOUR, TEXTURE and TASTE. The values that are formatted cursive, like *red-green-yellow* for the colour of the apple peel, refer to a conceptual space which specifies the corresponding region in colour space.

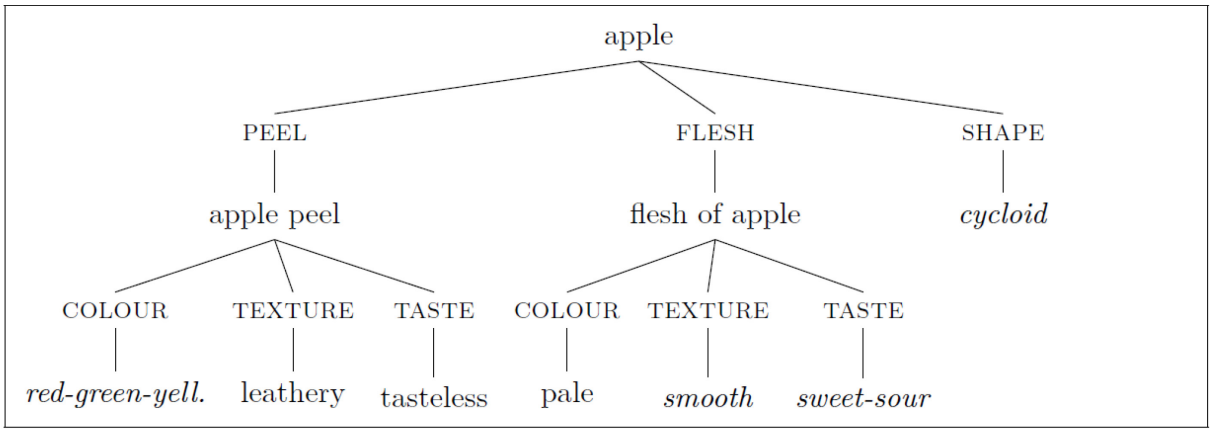

Figure 93: Criterion C as a (partial) frame of an apple. Figure 4 from (Strößner, 2020b, p. 694).

She notes that whenever attribute values are assumed to be quantified by joint probability distributions, conceptual spaces "are not only a *possible* extension of stochastic frames, but already an implicit part of them." (p. 700). She also notes that with the introduction of a similarity measure, albeit with different mathematical properties, the probabilistic prototype frame model as developed in this thesis has a connection to conceptual spaces, however "[w]hether it is possible to modify prototype frames in a way that makes them compatible to geometric representation and the question of how this should be done are matters for future research" (p. 702).

A very important property of conceptual spaces is the possibility to explain the naturalness of concepts with them. The research on conceptual spaces and prototypes has produced interesting results. Douven and Gärdenfors (2020) argue that the regions in conceptual spaces that represent natural concepts correspond to the optimal partitions of a similarity space. Douven (2019) presents interesting work showing that prototypes in colour space follow similarity constructed in this way.

While I do not think that exact specifications in terms of conceptual spaces would contribute to the predictive power of prototype frames, the explanation that properties constitute regions in conceptual spaces adds to the cognitive plausibility of probabilistic representations. They also aid in identifying possible values of attributes as the convex regions of a conceptual space.

9.2.3 **Mentalised frames**

With Leda Berio, I am working on incorporating mentalising, i.e., the representation of the representations of other people in order to successfully communicate, as an update-operation for frames. The idea behind this is illustrated in Figure 94. Imagine that I am walking with my dog Nala, who has blonde fur of medium length. I meet a person whom I do not know. My (CP1's) frame (left top) includes all the information that I know about my dog. The other person (CP2) can, upon seeing my dog, be assumed to have all visible information on her in their frame (left bottom), if the context specifies clear visibility and that CP2 is not blind. The attribute NAME has the value unknown (?) for them. When I talk to this person about my dog, I estimate the information that they have about my dog (right top) and compare it to my frame. This would be the mentalising operation. Correctly, in CP1's mentalised frame for CP2, the value of the NAME attribute is unknown (red) and thus CP1 knows that it cannot be taken as common ground in communication with CP2. In successful communication, like in this example, the known and estimated frames are identical.

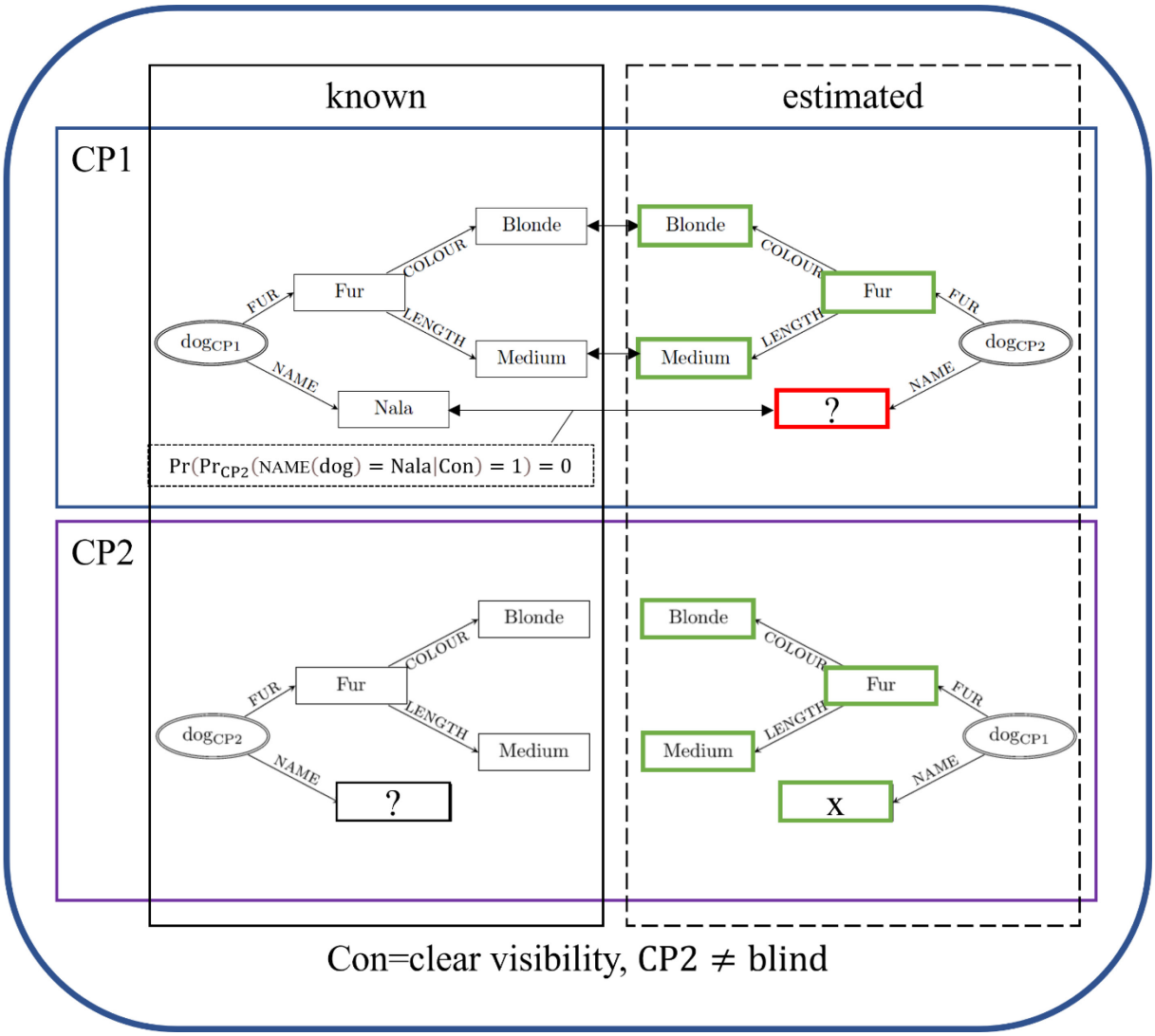

Figure 94: Mentalised frames for a specific dog by two communication partners (CPs) in a specified context Con.

This work has two interesting applications. First, to compare the own and mentalised frames collected from participants and investigate in how far they overlap. This is particularly interesting for concepts which can be assumed to have a large subjective component, like *love*, for which we plan to ask couples to estimate the frames of their partners and then compare the results. A second interesting application is to develop a frame that contains all the components required for successful communication in cases in which miscommunication must be avoided at all costs, like in doctor-patient-communication. For an example, consider Figure 95, which represents a communication situation of a doctor and a patient in which the doctor recommends the patient to follow a low-sodium diet. What the doctor assumes the patient to know is that salt affects the sodium intake, as it consists of sodium. The patient, however, has no value on the CHEMICAL FORMULA attribute of their *salt* frame. What the doctor could have done to avoid this misunderstanding is to communicate "Note that salt is sodium chloride", which would have led to the update of the patient's *salt* frame.

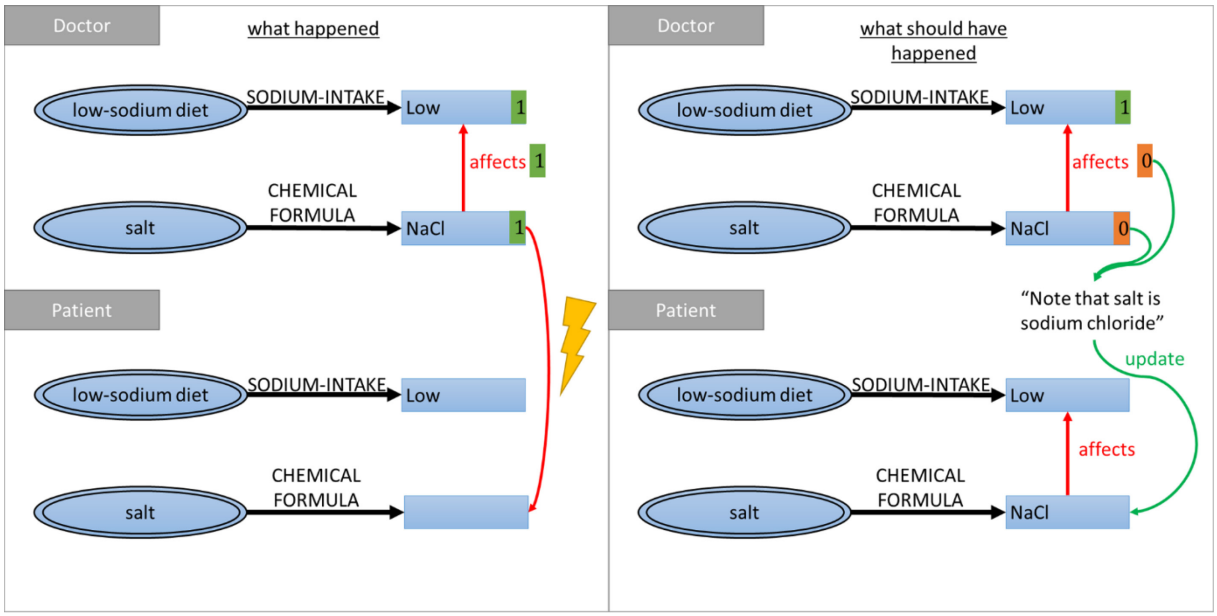

Figure 95: An example of unsuccessful (left) and successful (right) doctor-patient-communication.

Working with doctors to create frames that contain essential components of medical terms and sharing them with patients could aid doctor-patient-communication and contribute to a healthier society.

9.2.4 Extending the scope

The concepts investigated in this thesis, *fruit* and *vegetable*, are very interesting in terms of their formal properties, which for example allowed us to test Bayes' theorem for probability judgements. But they are (for most) not the most interesting concepts that exist. Contrary to the model of Smith et al., our model does not require contrast category data to calculate diagnosticity. To assess the generalisability of the model, it would be interesting to identify prototypes for more (interesting) concepts, which comprises on the one hand more "classical" prototype categories like *furniture* and *mammal*, and on the other hand more theoretical notions for which people in general have no deeper theoretical knowledge, which makes them no candidates for a cognitively plausible representation in terms of theory theory, like *intelligence* and *love*.

Furthermore, as mentioned in the context of conceptual combination (section 9.2.1), prototype frames could profit from embedding constraints which could mean a "boost" in typicality contribution for the combined presence of certain properties. For example, finding the properties *is red* and *is sweet* in one SC together could raise its typicality more than finding only one of those properties. In Malt and Smith (1984), this question was investigated and found to hold for "certain particularly salient or functional combinations" (p. 250), like *is sweet* and *tastes good*. As constraints become important in conceptual modification, it seems to be advantageous to collect them when specifying probabilistic prototypes, as ratings that estimate $Pr(P_k|P_l, C)$, i.e., "How probable is it that a fruit is red, given it is sweet?". Alternatively, constraints can be determined after the experiment by investigating the correlations between property probability ratings. Malt and Smith introduced the notion to the FRS by weighting the cooccurring properties with the number of times they cooccur in the category. Embedding an additional weight for cooccurring properties could also work for probabilistic prototype frames.

9.2.5 Exploring property sources

In section 4.2, I presented two alternatives to property generation data as the input for prototype representations: the large-scale word association project Small World of Words and properties derived from functional divisions in the brain presented in Binder et al. (2016). Both offer extensive amounts of interesting data, and it would be interesting to see how well they work as an input to the property probability experiments.

The advantage of the Small World of Words data is their incredibly high participant base in several languages which makes their data a lot more representative than generation studies with usually 30 participants per word. If a reliable procedure to translate the associations into properties, this would lead to a large amount of prototype frames that could be created with relatively little effort.

The advantage of the data from Binder et al. (2016) is that they define the structure of the frame with attributes that are associated with distinct processes in the brain. It lacks however the range of admissible values, which would have to be derived for the identified most important attributes. This could be done with a more precise property generation experiment in which participants are for example asked to generate all colours and actions associated with a category.

The procedure used in Hampton (1979) to identify category properties together with participants in interviews could, despite taking a lot of time, uncover the most useful set of properties.

I would like to investigate in how far these inputs make a difference to typicality prediction in a comparison with the parameters of precision of predictions and workload. If it can be shown that similar results are possible with data that is already available, this would facilitate future endeavours to identify cognitively plausible representations.

9.2.6 Exploring typicality measures

I found in the meta-analysis that typicality ratings are not intersubjectively stable in the medium typicality range. James Hampton proposed that it might be worthwhile to try to predict typicality orderings from different sources. One potential candidate are mean reaction times from categorisation decision experiments. It was shown already in Rosch (1975b) and many times after that it takes participants much less time to agree to the categorisation of typical members than untypical ones.

9.2.7 Methodological issues in collecting data for the research on concepts

In the course of this thesis, I collected and analysed several kinds of empirical data. The intersubjective stability of typicality ratings is the motivation for research on prototypes in general as well as for the research presented here. In the typicality meta-analysis and in my own collection of typicality data, I showed that it is necessary to not only report standard deviations, which is not consistently done in the literature, but also to examine the rating distribution for each subcategory included in the experiment in order to identify whether a high SD is due to a skewed distribution or due to a multi-modal distribution. In the latter case, it is not possible to consider the ratings for this SC to be intersubjectively stable.

The other datatype I collected are property probability ratings, for which I found a high variance and a trend to use extreme ratings, whether it made sense for the property in question or not. Future researchers who are interested in subjective frequency estimations should write a precise explanation that avoids ambiguity and prevents participants from rating their own uncertainty,

the possibility of this property for the object in question and the grade to which this property applies. For my thesis, I decided to exclude all pairs that had a high variance and to focus on those for which intersubjective agreement could be seen. This meant unfortunately that almost 50% of pairs were excluded.

10 APPENDIX

10.1 Participants

Table 83: Age, gender, English dialect, country of residence, level of education and botany knowledge of participants in experiment 3

		experiment 1	experiment 2	experiment 3	
				fruit	vegetable
Gender	male	11	14	10	9
	female	19	16	20	21
	other	0	0	0	2
Age	[18,30)	17	14	15	17
	[30,40)	8	8	6	5
	[40,50)	4	4	4	7
	[50,60)	1	3	3	1
	[60,70)		1	2	2
Country of residence	UK	20	19	23	18
	USA	7	7	5	7
	Canada	3	1	2	4
	Ireland	0	2	0	0
	Australia	0	1	0	2
	Spain	0	0	0	1
English dialect	British	18	21	25	19
	American	10	9	4	9
	Others	2	0	1	4
Education	Less than high school	1	0	0	1
	High school graduate	2	3	6	8
	Some college	7	8	6	9
	2 year degree	4	3	2	1
	4 year degree	2	2	0	3
	Professional degree	1	1	4	0
	Bachelor	8	8	7	6
	Master	2	5	2	4
	Doctorate	3	0	3	0
Botany knowledge	Yes	1	4	5	7
	No	29	26	25	25

10.2 **Experiment 1: Property stimuli**

Table 84: Properties used as stimuli in experiment 1 with category production frequencies (CPFs).

property	Devereux et al. formulation	McRae et al. formulation	McRae et al. CPF fruit	McRae et al. CPF vegetables	Devereux et al. CPF fruit	Devereux et al. CPF vegetables
is black	is black		0	0	5	1
is brown	is brown		0	0	7	5
is green	is green	is_green	6	12	10	14
is orange	is orange	is_orange	6	2	8	2
is pink	is pink		0	0	4	1
is purple	is purple	is_purple	3	2	5	3
is red	is red	is_red	7	2	12	4
is white	is white	is_white	2	5	3	9
is yellow	is yellow	is_yellow	8	3	9	3
is eaten as dessert	is eaten as dessert		0	0	5	0
is eaten in salad	is eaten in salad	eaten_in_salads	0	11	0	6
is for soup	is for soup	eaten_in_soups	0	5	0	3
is made into juice	is made into juice	used_for_juice	7	0	4	0
is eaten in pies		eaten_in_pies	4	0	0	0
is baked	is baked		0	0	0	3
is boiled	is boiled	eaten_by_cooking	0	13	0	10
is chopped	is chopped		0	0	0	3
is dried	is dried		0	0	6	0
is eaten raw	is eaten raw	eaten_raw	1	3	0	5
is fried	is fried		0	0	0	3
is peeled	is peeled	eaten_by_peeling	4	0	4	3
is roasted	is roasted		0	0	0	3
grows in the ground	does grow in ground	grows_in_the_ground	0	7	2	16
grows on bushes	does grow on bushes	grows_on_bushes	4	0	3	0
grows on plants	does grow on plants		0	0	1	4
grows on trees	does grow on trees	grows_on_trees	17	0	18	0
grows on the ground		grows_on_the_ground	0	3	0	0
grows on vines		grows_on_vines	1	1	0	0
grows underground		grows_underground	0	4	0	0
is white inside	has white flesh	is_white_inside	1	4	1	4
has yellow flesh	has yellow flesh		0	0	3	0
has a core	has a core		0	0	4	0
has a heart	has a heart		0	0	0	3
has a pit/stone	has a stone	has_a_pit	5	0	8	0
has flesh	has flesh		0	0	18	6
has layers	has layers		0	0	0	4
has pips/seeds	has pips_seeds	has_seeds	13	4	15	4
has sections	has segments	has_sections	3	0	6	0
has an inside		has_an_inside	5	4	0	0
has green leaves	has green leaves		0	0	1	3
has leaves	has leaves	has_leaves	1	8	2	12
has sugar	has sugar		0	0	3	0

Table 84 [cont.]: Properties used as stimuli in experiment 1 with CPFs.

property	Devereux et al. formulation	McRae et al. formulation	McRae et al. CPF fruit	McRae et al. CPF vegetables	Devereux et al. CPF fruit	Devereux et al. CPF vegetables
has vitamins	has vitamins	has_vitamin_C	4	0	9	3
is healthy	is healthy		0	0	20	18
made of carbohydrate/starch	made of carbohydrate_starch		0	0	0	4
is nutritious		is_nutritious	3	10	0	0
has a stalk/stem	has a stalk_stem		0	0	7	11
has flowers	has flowers		0	0	0	3
has roots	has roots		0	0	0	5
has stalks		has_stalks	0	3	0	0
has pith	has pith		0	0	5	0
is a bulb	is a bulb		0	0	0	3
is round	is circular_round	is_round	15	7	22	10
is long	is long	is_long	1	5	1	8
is thin	is thin		0	0	0	7
is big	is big_large		0	0	3	1
is small	is small	is_small	13	3	20	6
has hard/tough skin	has hard_tough skin		0	0	5	0
has thick skin	has thick skin		0	0	5	0
is furry	is furry		0	0	3	0
is wrinkly	is wrinkly		0	0	5	0
has peel	has skin_peel	has_peel	6	0	25	9
has skin	has skin_peel	has_skin	12	3	25	9
has zest	has zest		0	0	3	0
is smelly	does smell_is smelly		0	0	0	4
has a strong flavour	has a strong flavour		0	0	0	3
tastes bad	is disgusting_taste bad		0	0	2	7
is sour	is sour_sharp_acidic	tastes_sour	6	1	10	0
is sweet	is sweet	tastes_sweet	20	3	25	3
is tasteless/bland	is tasteless_bland		0	0	0	4
is tasty	is tasty	tastes_good	13	7	27	22
tastes tart		tastes_tart	3	0	0	0
has soft flesh	has soft flesh		0	0	3	1
is crunchy	is crunchy	is_crunchy	1	7	1	7
is hard	is hard		0	0	4	5
is juicy	is juicy	is_juicy	18	2	23	1
is soft	is soft	is_soft	4	0	13	4
is squashy/squidgy/squishy	is squashy_squidgy_squishy		0	0	3	0
is watery	made of water_is watery		0	0	2	5
is used for baking	is used for baking		0	0	3	0
is used in cooking	is used in cooking		0	0	9	20
is eaten in summer	is associated with summer	eaten_in_summer	5	0	4	1
grows in hot countries	does grow in hot countries	grows_in_warm_climates	4	0	12	0
is tropical	is tropical		0	0	3	0
grows in gardens		grows_in_gardens	0	15	0	0

10.3 **Experiment 1: Summary statistics for all properties**

Table 85: Summary statistics for all properties from experiment 1.

property	mean	SD	median	IQR	range
is eaten as dessert	4.0	1.5	5.0	1.0	5.0
has zest	3.3	1.7	4.0	2.8	5.0
is sweet	2.7	1.7	3.0	2.0	5.0
is juicy	2.7	1.7	3.0	3.0	5.0
tastes tart	2.5	1.9	3.0	3.0	7.0
has a pit/stone	2.5	2.2	2.5	4.5	8.0
has sugar	2.4	1.7	2.5	3.0	5.0
is tropical	2.2	1.8	2.5	3.8	5.0
grows on trees	2.2	1.8	2.5	3.8	5.0
has pips/seeds	2.1	2.0	2.0	4.0	5.0
grows on bushes	2.1	2.0	2.5	3.8	7.0
is made into juice	2.1	1.8	2.0	4.0	5.0
is sour	1.9	2.5	2.5	4.0	10.0
has pith	1.8	1.9	1.0	3.0	5.0
is pink	1.6	1.9	1.0	3.0	5.0
is used for baking	1.6	2.5	0.0	4.0	9.0
grows on vines	1.6	2.0	1.0	3.0	7.0
is furry	1.5	1.7	1.0	3.0	6.0
is dried	1.4	2.0	0.0	3.0	7.0
has a core	1.3	2.1	1.0	3.0	10.0
is watery	1.3	1.7	0.0	2.0	5.0
has a peel	1.3	2.0	0.0	2.0	7.0
is soft	1.2	1.5	0.0	2.8	5.0
grows on plants	1.1	1.8	0.0	2.8	7.0
is eaten in pies	1.1	2.3	0.0	3.0	10.0
is eaten in summer	1.0	1.3	0.0	2.0	4.0
has sections	1.0	2.0	0.0	2.8	8.0
is squashy/squidgy/squishy	0.8	2.2	0.0	2.0	10.0
is orange	0.8	1.4	0.0	1.8	5.0
has soft flesh	0.8	1.8	0.0	1.8	10.0
grows in hot countries	0.8	1.8	0.0	1.8	9.0
has flesh	0.7	1.4	0.0	1.0	6.0
is red	0.7	1.2	0.0	1.0	6.0
is eaten raw	0.7	1.1	0.0	1.0	4.0
is round	0.6	1.4	0.0	0.8	7.0
is tasty	0.6	1.4	0.0	0.0	5.0
is yellow inside	0.5	1.2	0.0	0.0	6.0
has a strong flavour	0.5	1.5	0.0	0.8	7.0
is small	0.4	0.7	0.0	1.0	3.0
has flowers	0.4	2.1	0.0	1.0	10.0
is wrinkly	0.4	1.4	0.0	0.0	7.0
is yellow	0.2	0.6	0.0	0.0	2.0
has vitamins	0.2	0.8	0.0	0.0	4.0
has an inside	0.1	0.7	0.0	0.0	5.0
is peeled	0.0	1.6	0.0	0.0	10.0
has skin	0.0	1.2	0.0	0.0	8.0
is nutritious	-0.1	0.9	0.0	0.0	6.0
has thick skin	-0.2	1.8	0.0	0.0	9.0
is crunchy	-0.2	0.8	0.0	0.0	4.0
grows in gardens	-0.3	0.7	0.0	0.0	3.0
is black	-0.4	1.8	0.0	0.8	9.0
is healthy	-0.4	1.1	0.0	0.0	5.0
is purple	-0.4	1.0	0.0	0.8	5.0
is white inside	-0.4	1.0	0.0	0.0	5.0
tastes bad	-0.6	1.1	0.0	1.0	5.0
is hard	-0.6	1.3	0.0	1.0	7.0
is chopped	-0.6	1.2	0.0	0.8	4.0
is big	-0.7	1.2	0.0	1.0	5.0
is white	-0.7	1.4	0.0	0.8	5.0
has green leaves	-0.7	1.5	0.0	0.0	6.0
is smelly	-0.7	1.5	0.0	1.8	8.0
is thin	-0.7	1.2	0.0	1.8	3.0
has leaves	-0.8	1.3	0.0	1.8	4.0
has hard/tough skin	-0.8	1.4	0.0	2.0	6.0
is long	-0.9	1.3	0.0	2.0	5.0
has a stalk/stem	-0.9	2.2	0.0	2.0	9.0
is baked	-0.9	2.2	0.0	2.0	10.0
is used in cooking	-1.0	1.5	0.0	2.0	5.0
has stalks	-1.2	1.7	0.0	2.0	6.0
has a heart	-1.3	1.7	0.0	3.0	5.0
is green	-1.4	1.8	0.0	2.8	5.0
is brown	-1.5	1.8	-0.5	3.0	5.0
is eaten in salad	-1.5	2.1	-1.0	3.0	8.0
has layers	-1.6	2.0	-1.0	3.0	7.0
grows on the ground	-1.7	1.9	-1.0	3.0	5.0
is tasteless/bland	-1.7	1.7	-1.5	3.0	5.0
is a bulb	-2.1	2.3	-2.0	4.0	8.0
is made of carbohydrate/starch	-2.5	1.8	-3.0	3.8	5.0
is boiled	-2.6	1.9	-3.0	2.8	5.0
grows in the ground	-2.7	2.0	-3.0	4.5	5.0
has roots	-2.8	1.8	-3.0	2.0	5.0
is fried	-2.9	1.8	-3.0	3.5	5.0
is roasted	-3.2	1.6	-4.0	1.8	5.0
grows underground	-3.3	2.0	-4.0	3.0	5.0
is for soup	-3.7	1.6	-4.0	2.0	5.0

10.4 **Experiment 2: Summary statistics**

Table 86: Means and SDs of property probability ratings for fruit and vegetables from experiment 2 and difference of means.

property	mean fruit	SD fruit	mean vegetables	SD vegetables	diff
is made into juice	4.7	0.6	2.7	1.5	1.9
has sugar	4.6	0.8	2.1	1.4	2.5
is eaten in summer	4.4	0.8	3.3	1.6	1.1
has pips/seeds	4.4	0.8	2.2	1.8	2.2
is sweet	4.4	0.7	1.8	1.4	2.6
is juicy	4.4	0.7	1.7	1.3	2.7
grows on trees	4.3	0.8	1.6	1.5	2.7
is eaten as dessert	4.2	1.1	0.8	1.1	3.4
is tropical	3.9	0.9	1.7	1.2	2.3
has a peel	3.9	0.9	2.9	1.6	1.0
grows on bushes	3.9	1.0	1.9	1.6	2.0
has a core	3.8	0.9	2.1	1.3	1.7
is soft	3.7	1.1	2.3	1.3	1.4
has a pit/stone	3.7	1.0	1.1	1.1	2.6
is watery	3.6	1.5	2.3	1.5	1.3
is eaten in pies	3.6	1.4	2.7	1.7	0.9
is dried	3.6	1.3	1.9	1.6	1.8
is used for baking	3.6	1.1	2.2	1.4	1.5
is used in cooking	3.6	1.4	4.8	0.5	1.2
grows on plants	3.6	1.3	3.4	1.6	0.2
has sections	3.5	1.1	2.2	1.4	1.3
has zest	3.5	1.0	1.0	1.3	2.4
grows on vines	3.2	1.2	2.1	1.5	1.2
has pith	3.1	1.4	1.3	1.4	1.8
tastes tart	3.0	1.4	1.4	1.2	1.6
is pink	2.9	1.4	0.9	1.2	1.9
is sour	2.7	1.4	1.5	1.3	1.3
is furry	2.7	1.4	0.9	0.9	1.8
is green	2.6	1.4	3.9	1.3	1.3
is eaten in salad	2.5	1.5	3.9	1.4	1.4
is made of carbohydrate/starch	2.5	1.8	3.6	1.4	1.1
has layers	2.4	1.7	3.2	1.2	0.8
grows on the ground	2.3	1.7	3.8	0.9	1.5
has stalks	2.2	1.5	3.3	1.2	1.2
grows in the ground	1.8	1.5	4.1	0.8	2.3
has roots	1.7	2.0	4.1	1.0	2.3
is brown	1.6	1.0	3.0	1.2	1.4
is fried	1.6	1.5	3.3	1.4	1.7
is boiled	1.5	1.4	4.1	0.8	2.5
is roasted	1.3	1.3	4.1	1.3	2.9
is a bulb	1.1	1.0	2.5	1.6	1.4
is for soup	1.0	1.3	4.4	0.8	3.4
has a heart	0.9	1.4	1.5	1.8	0.7
is tasteless/bland	0.9	1.3	2.1	1.2	1.2
grows underground	0.6	0.7	3.6	1.2	3.0

10.5 **Experiment 3: Summary statistics for typicality ratings**

Table 87: Mean and SD typicality ratings before and after cleaning and amount of "not a category member" ratings (NM) for a) fruit and b) vegetable.

a)

subcategory	mean	SD	mean before cleaning	SD before cleaning	Median	IQR	IQR after cleaning	NM
apple	1.0	0.2	1.4	1.2	1.0	0.0	0.0	0
strawberry	1.2	0.5	1.6	1.3	1.0	0.8	0.0	0
banana	1.2	0.4	1.6	1.4	1.0	0.8	0.0	0
peach	1.2	0.5	1.6	1.4	1.0	1.0	0.0	0
grape	1.2	0.7	1.7	1.5	1.0	0.8	0.0	0
pineapple	1.2	0.6	1.8	1.7	1.0	1.0	0.0	0
mango	1.4	0.8	1.7	1.4	1.0	1.0	0.0	0
blueberry	1.4	0.8	1.9	1.6	1.0	1.0	0.0	0
passion fruit	1.4	0.8	2.0	1.7	1.0	1.0	1.0	0
blackberry	1.5	0.9	2.0	1.6	1.0	1.8	1.0	0
watermelon	1.5	1.1	2.0	1.6	1.0	1.8	0.0	1
plum	1.6	0.9	2.0	1.4	1.0	2.0	1.0	0
pomegranate	1.6	1.2	1.8	1.2	1.0	1.8	1.0	0
lime	2.2	1.7	2.5	2.0	2.0	3.0	2.0	1
papaya	2.4	1.9	2.6	2.0	2.0	3.0	2.0	0
prune	2.9	1.7	3.1	1.7	3.0	2.0	3.0	1
fig	3.3	2.0	3.6	2.1	4.0	4.0	3.0	3
rhubarb	3.7	2.5	3.8	2.4	4.0	4.8	5.0	7
avocado	4.4	1.9	4.4	2.0	4.5	2.8	2.0	4
coconut	4.5	1.8	4.7	1.9	4.0	3.5	3.0	6
tomato	4.7	1.9	4.3	2.1	4.0	3.0	4.0	3
pumpkin	5.2	1.8	4.9	2.0	5.0	3.8	3.0	7

b)

subcategory	mean	SD	mean before cleaning	SD before cleaning	median	IQR	IQR after cleaning	NM
broccoli	1.2	0.6	1.7	1.6	1.0	0.0	0.0	0
cauliflower	1.3	0.8	1.8	1.5	1.0	1.0	0.0	0
carrot	1.3	0.8	1.5	1.2	1.0	0.3	0.0	0
pea	1.6	0.9	2.1	1.6	1.0	1.3	1.0	2
spinach	2.1	1.8	2.4	1.9	1.5	2.3	1.0	2
lettuce	2.1	1.5	2.5	2.0	1.5	2.3	2.0	1
zucchini	2.2	1.9	2.5	2.2	1.0	2.3	2.0	3
onion	2.3	1.3	2.4	1.6	2.0	2.0	2.0	0
green onion	2.4	1.7	2.4	1.6	2.0	2.3	2.0	0
sweet potato	2.5	1.4	2.7	1.7	2.0	3.0	3.0	0
corn	2.6	1.7	3.0	1.9	2.5	3.0	3.0	2
eggplant	2.6	1.9	2.6	2.0	1.5	3.0	3.0	4
potato	2.7	1.8	2.7	2.0	2.0	3.0	3.0	0
radish	3.0	2.0	3.0	1.9	2.5	3.3	4.0	0
mushroom	3.7	2.2	3.6	2.2	3.5	4.3	5.0	12
pumpkin	3.8	1.9	3.9	2.1	4.0	3.3	2.0	7
tomato	4.3	2.3	4.1	2.4	4.0	6.0	4.0	17
garlic	4.4	2.0	4.3	2.0	4.0	3.0	3.0	5
pickle	5.0	2.1	4.8	2.1	5.0	4.3	3.0	8
avocado	5.0	1.9	4.7	2.2	5.0	4.0	3.0	16
parsley	5.4	1.9	5.1	2.1	6.0	3.3	3.0	13
rhubarb	5.4	1.9	5.1	2.0	5.0	3.3	3.0	12

10.6 **Experiment 3: Summary statistics for probability ratings**

Table 88: Summary statistics for all fruit SC property probability ratings with mean and SD in brackets.

subcategory	will be eaten in soup	will be roasted	grew in the ground	has roots	will be boiled	will be made into juice	has sugar	has pips/seeds	is sweet	is juicy	grew on a tree	will be eaten as dessert	tastes tart	has a heart	will be used in cooking	grew on a plant
apple	11	42.2	5.1	11.4	46.3	82.2	85.3	93.9	84.4	78.4	95.6	76.7	49	12.9	68.8	37.4
	(14.7)	(34.4)	(10.7)	(22.9)	(31.9)	(20.9)	(20.3)	(11.8)	(15.1)	(20.8)	(9.6)	(22.3)	(28)	(26.4)	(25.1)	(44.4)
strawberry	8.4	14.1	21.5	20.9	24.1	60.4	90.7	78.7	89.6	84.8	7	86.9	37.2	9	51.9	95.3
	(15.1)	(15.8)	(35.6)	(31.6)	(24.2)	(32.9)	(11.4)	(31.3)	(12)	(17.7)	(19)	(15.2)	(29.9)	(17.1)	(29.3)	(12.1)
banana	8.5	33.3	8.3	10	8	42.7	82.1	32.2	73.6	25.6	78.7	76.6	21.4	5.5	49.9	48.7
	(13)	(34.4)	(18.9)	(24.2)	(13.4)	(35.6)	(22.8)	(35.9)	(27.6)	(29)	(36.3)	(24.1)	(25.6)	(11.4)	(31)	(45.8)
peach	16.1	29.3	5.2	10.7	24	64.6	83	69.7	87.5	84.4	81.8	79	42	11.1	51.7	50.8
	(26.5)	(28.3)	(8.3)	(23.3)	(25.7)	(28.8)	(19.6)	(35.6)	(13.6)	(15.4)	(31.5)	(22.2)	(33)	(22)	(29.7)	(41.4)
grape	8.4	10.2	14.1	16.1	16.6	77.5	86.1	75.1	83.9	87.1	29.8	73.1	36.9	4.9	43.8	85.3
	(15.1)	(17)	(23.2)	(26.6)	(24.6)	(20.8)	(21)	(20.8)	(16.6)	(15.3)	(40.7)	(26)	(25.8)	(7.9)	(28.1)	(28.5)
mango	11.6	28.4	8.3	16.7	23.6	73.3	82.8	58.2	84.1	81.6	85.3	81.6	32.9	7.6	57.2	53.5
	(18.6)	(34.5)	(14.9)	(26.5)	(30.5)	(26.2)	(17.4)	(38.1)	(16.7)	(21.2)	(25.3)	(19.8)	(25)	(14.8)	(30.7)	(40.9)
pineapple	14.9	42.6	24.1	20	18.5	81.9	86.6	33.4	80.7	89.5	59.1	77.4	42.7	18.1	63	62.9
	(18.5)	(35.4)	(34.6)	(32)	(22.5)	(17.3)	(16.1)	(36.8)	(18)	(11)	(42.1)	(20.1)	(33.5)	(30.5)	(30.3)	(39.7)
pomegranate	19.7	22.2	15.5	16.4	17.8	69.5	76.9	90.2	74	75.7	74.5	68.3	46.7	9.8	45.2	50.9
	(24.3)	(21)	(22.5)	(25.5)	(20)	(23.3)	(25.6)	(15.6)	(18.9)	(22.1)	(32.1)	(25.4)	(31.9)	(18)	(31.6)	(40.3)
blueberry	10.1	9.5	10.3	16.8	26.5	70.3	75.8	55.9	74.3	76.8	24.7	78.4	56.5	5	47.8	89.9
	(16.4)	(17)	(20.8)	(29.4)	(31.5)	(25.2)	(26.3)	(37.3)	(22.3)	(23.8)	(37.1)	(22.9)	(26.8)	(9.2)	(35.7)	(21.3)
plum	14.4	30	16	16.7	38.3	60.4	81.5	75.8	76.6	74.8	86.5	71.9	47.4	9.3	51.1	40.5
	(20.9)	(29.3)	(28.9)	(29)	(32.4)	(34.3)	(18.8)	(32.8)	(21)	(24.3)	(25.7)	(23)	(28.1)	(20.5)	(29.3)	(42)
blackberry	11.1	14.8	6.5	10.8	34.6	65.5	82.2	68.7	69.4	78.1	24.2	76.9	59.1	8.7	57.2	85.7
	(17.1)	(17.9)	(15.2)	(21.4)	(31.6)	(30.5)	(18.1)	(31.7)	(26.6)	(19.5)	(36.8)	(25.4)	(26.9)	(21.2)	(31.4)	(27.4)
watermelon	13.1	17	42.6	26	9.4	64.7	81.5	95.3	84.3	92.1	22.1	75.1	22.1	6.5	24.4	73.6
	(19.1)	(22.2)	(40.4)	(34.1)	(11.7)	(30.1)	(19.7)	(8.4)	(21.1)	(10.7)	(31.6)	(28.2)	(25.6)	(10.1)	(22.5)	(37)
passion fruit	14.6	21	13.6	19	21.2	68.7	78.1	86.5	70.8	78	67.7	74.6	52.9	10	42.8	61.5
	(19.4)	(27)	(22.2)	(31)	(22.7)	(28.2)	(24.8)	(23)	(30)	(23)	(34.5)	(23.9)	(33.4)	(18.3)	(30.5)	(37.6)
lime	13.3	13.2	6.9	11.1	13.7	68.9	58.5	74.8	30	86.2	87.3	45.6	83.9	6.6	67.1	42.8
	(17.4)	(17.4)	(17.3)	(22.7)	(19.5)	(31.8)	(30.8)	(32.7)	(28.5)	(20.9)	(27.8)	(34.2)	(20.4)	(11.6)	(32.2)	(45.2)
papaya	23.5	30.8	20.5	19.9	27.7	58.1	74.5	68.4	73.8	65.8	71.7	57.7	38.6	13.1	40.8	56.5
	(28.6)	(25.6)	(26.5)	(28)	(30.1)	(32.1)	(25.4)	(30.5)	(19.6)	(24.2)	(31.7)	(31.7)	(28.9)	(19.3)	(32.9)	(41.1)
prune	14.3	27	15.1	17	33.6	70.1	76.3	65.9	62.8	48.3	69.2	68.8	48	7	61.3	61.6
	(17)	(31.2)	(19.8)	(26.3)	(29.5)	(22.4)	(20.2)	(32.1)	(29.6)	(27)	(36.9)	(28.1)	(30.5)	(12.4)	(27.1)	(37.1)
fig	17.8	46.7	12.5	13.9	32	49.5	73.6	77.8	67.8	48.6	79.6	68.8	42.9	8.3	64.3	53.8
	(24.6)	(28.7)	(19.5)	(22.3)	(29.9)	(31.5)	(23.6)	(27.8)	(22.7)	(27.7)	(30.6)	(26.2)	(27)	(14.7)	(25.8)	(41.7)
rhubarb	27.7	41.5	67.6	59.2	63.1	38	60.1	20.7	45.5	43.3	8.8	67.8	73.7	8.9	78	57
	(31.3)	(31.5)	(41.8)	(37.2)	(29.3)	(30.6)	(31.9)	(28.4)	(26.8)	(28.6)	(14.6)	(31.7)	(24)	(19.6)	(22.7)	(44.2)
tomato	77.3	66	20.9	25.2	49.3	67.6	56.9	88.2	48.9	84.2	9.7	15.7	43.6	6.8	85.9	91.2
	(27.1)	(31.4)	(30.5)	(36)	(37.6)	(32.2)	(31.1)	(22.3)	(28.4)	(19.1)	(20.9)	(22.5)	(31.2)	(13.2)	(22.1)	(23)
avocado	30.6	22.2	15.6	13.5	14.5	28.3	44.9	65.2	31	30	68.7	29	26.1	14.5	58.3	63.2
	(31.2)	(25.5)	(27.8)	(26)	(16.2)	(33.2)	(28.6)	(37.8)	(26.7)	(30.2)	(39.4)	(30.5)	(30.9)	(25.6)	(32.3)	(41.4)
coconut	24	38.6	7.5	9.3	32.8	61.7	62.2	18.9	60.1	51.5	94.2	68.7	21.6	11.8	71.7	47.2
	(28.3)	(36.2)	(12.9)	(16.9)	(33.2)	(36.6)	(34.2)	(28.7)	(30.5)	(38.9)	(17.3)	(25.7)	(25.4)	(25.5)	(25.3)	(45.8)
pumpkin	73.7	69.6	61.7	44.4	46.5	41.5	56	90	44.6	44.4	7.8	54	36.1	10.7	78.5	65.7
	(23.7)	(29.5)	(42.4)	(36.1)	(30.4)	(34.8)	(24.7)	(16.5)	(22.2)	(29.1)	(15.7)	(30.4)	(28.8)	(19.4)	(25.5)	(40.9)

Table 89: Summary statistics for all vegetable SC property probability ratings with mean and SD in brackets.

subcategory	will be eaten in soup	will be roasted	grew in the ground	has roots	will be boiled	will be made into juice	has sugar	has pips/seeds	is sweet	is juicy	grew on a tree	will be eaten as dessert	tastes tart	has a heart	will be used in cooking	grew on a plant
carrot	73.2	65.5	96.1	73.6	76.1	62.8	51.3	10.3	53.1	32.8	4.5	28.9	15	2.7	86.2	31.9
	(24.5)	(27.7)	(9.9)	(32)	(23.3)	(27)	(36.3)	(24.6)	(26.1)	(27.6)	(7.5)	(28.1)	(18.8)	(4.3)	(18.6)	(39.9)
broccoli	60.7	51.4	55.7	49.1	73.4	27.9	33.4	12.8	15.2	15.7	5.9	3.8	13.1	7.9	86.7	49.1
	(33.3)	(31)	(41.9)	(37.9)	(26.1)	(29.4)	(39)	(24.4)	(18.9)	(24)	(11.2)	(4.7)	(19.2)	(20.7)	(16.6)	(41.4)
cauliflower	60.9	58.5	70.5	57.2	70	13.8	34	8.2	22.7	14.2	9.6	8.8	20.7	13.9	86.4	45.6
	(28.4)	(31)	(37.7)	(38.5)	(27.6)	(19)	(37.4)	(19.1)	(24.9)	(17.4)	(19.9)	(9.2)	(29.1)	(28.5)	(20.6)	(39.2)
pea	62.9	29.3	24.9	22.9	71.4	4.6	84.3	13.4	49.5	19.7	51.4	30.3	10.1	6.3	9.5	85.4
	(28.7)	(29.4)	(33.7)	(33.2)	(28.8)	(7.9)	(27.3)	(19.4)	(39.3)	(32.5)	(25.7)	(29)	(17.5)	(8.4)	(12.5)	(18.2)
green onion	63.1	45.9	80.9	75.9	30.9	13.6	35.5	6.9	20.8	26.2	4.8	4.6	27.7	11.8	80.8	46.9
	(32.1)	(33.4)	(27.6)	(32)	(32.1)	(21.1)	(39)	(12.9)	(20.3)	(26.4)	(6.5)	(6.4)	(28.6)	(26)	(22.2)	(42.6)
spinach	59.7	24.8	54.5	38.1	52.7	43.9	34.4	7.7	22.7	20.6	8.4	6.8	20	8.2	80	56.4
	(31.3)	(29.8)	(40.8)	(40.8)	(31.9)	(33.7)	(38.1)	(14.4)	(22.8)	(21.6)	(16.7)	(10.1)	(21.5)	(18.9)	(23.6)	(40.7)
onion	69.4	63.8	88.9	72.6	41.9	10.8	43	6.9	31.3	35.9	8.1	6.3	31.4	11.2	88.9	37.4
	(26.9)	(29.6)	(25.4)	(35.9)	(27.6)	(17)	(40.5)	(11.1)	(27.8)	(31.5)	(19)	(9.6)	(34.1)	(25)	(16.5)	(43.4)
lettuce	22	17.1	67.2	52.3	14.5	13.8	35.9	9.1	30.2	27.2	5.7	5.7	12.4	33.5	45.4	46.9
	(28.6)	(28.4)	(36.1)	(39.5)	(18.2)	(21.4)	(38.7)	(20.3)	(28.5)	(28.5)	(12.6)	(6.7)	(18.6)	(39)	(38.3)	(42)
zucchini	48.5	65.9	51.3	32	40.3	16.4	34.2	59.8	26.4	35.6	14.8	14.9	23.5	7.1	76.9	64
	(29)	(30.9)	(40)	(34.9)	(30)	(24.6)	(36.2)	(40.8)	(25.6)	(33.3)	(25.6)	(18.4)	(26.7)	(12.8)	(29.1)	(38.6)
eggplant	43.8	68.5	41	39.9	42.5	9.2	35.4	54.7	27.4	31.1	20.8	11.1	24.4	8.3	82.1	67.7
	(30.3)	(31.5)	(37.5)	(39.4)	(29.4)	(11.4)	(36.1)	(39.8)	(27.5)	(29.6)	(28.9)	(14.7)	(29.2)	(14.9)	(23.6)	(36)
sweet potato	62.8	77.6	89.3	66.4	60.3	19.9	62.2	4.8	57.2	14.1	5.1	28.3	18.2	2.3	89.3	50.1
	(29.1)	(22.7)	(21.6)	(35.9)	(28.5)	(26.2)	(34.9)	(7.5)	(32.4)	(17.4)	(8.5)	(25.6)	(21.6)	(3.1)	(16.6)	(44.6)
potato	67.9	77.1	92.7	63	78.1	8.5	38.9	4	21.8	16.4	5.3	7.2	15.6	6.7	89.7	38.3
	(26)	(25.2)	(17.3)	(37.2)	(24.1)	(11.5)	(36)	(5.1)	(22.8)	(20)	(10.2)	(9.8)	(27.4)	(13.9)	(19.3)	(39.9)
radish	37.8	34.8	78	73.7	20.5	16	36.4	11.6	17.7	22.8	9.3	6.7	36.3	8.7	55.5	47.5
	(27.4)	(29.7)	(33.2)	(28.8)	(22.5)	(20.5)	(38.4)	(16.3)	(19.1)	(25.4)	(18)	(8.4)	(33.3)	(14.6)	(35.1)	(41)
corn	54.6	56	39.2	28.4	66.3	13.2	59.1	30.4	65	49.2	11.5	11.3	14.6	11.6	84.4	82.3
	(30.6)	(32.3)	(41.2)	(35)	(28.8)	(16.3)	(36.2)	(38.9)	(27)	(30.2)	(23.4)	(14.5)	(20.9)	(21.9)	(20.2)	(31.9)
mushroom	66.5	58.2	75.6	34.7	30.8	8.5	32.9	4.8	14.3	21.7	11.3	6.1	16.7	2.7	81.7	18.8
	(28.6)	(32.2)	(30.6)	(41.3)	(30.2)	(10.4)	(37.4)	(7.4)	(19.6)	(29.6)	(18.8)	(9)	(23.9)	(4.4)	(19.7)	(28.4)
pumpkin	68	64.1	55.3	42.3	49.1	40.6	51.5	86.8	51.7	42.9	4.3	61.2	29.6	6.5	78.9	57.3
	(29.3)	(29)	(40.3)	(40.2)	(30.6)	(35.6)	(38.3)	(26.8)	(31.7)	(33.2)	(6.3)	(30.7)	(30)	(11.5)	(22.9)	(41)
tomato	64.1	63.6	31	25.6	39.8	64.5	57.9	86.1	58.8	81.9	17.3	13.4	39.7	7.9	82.1	91.1
	(29.5)	(29.4)	(35.8)	(33.7)	(31)	(27.5)	(34.5)	(24.7)	(26.3)	(21.2)	(27.2)	(17.5)	(31.8)	(14.1)	(18.1)	(19.5)
garlic	61.8	68.2	75.1	58.4	24.4	9.1	41.6	15.4	31.3	8.4	14.2	18.4	4.7	3.6	34.5	93.3
	(29.1)	(28.1)	(35.4)	(41.7)	(26.1)	(20.2)	(40.2)	(19.8)	(36.8)	(13.1)	(15.6)	(21.6)	(7.2)	(4.9)	(33.6)	(11.3)
avocado	27	30.3	22.5	21.9	12.8	28	45.8	79.4	31.6	21.6	71.3	28.3	12.7	13.8	64.1	62.1
	(26.9)	(31.3)	(30.1)	(31.9)	(13.7)	(31.2)	(35.2)	(34.6)	(30.4)	(23.5)	(35.7)	(28.7)	(18.4)	(25.6)	(30.6)	(39.4)
pickle	19.1	21.2	29.7	12.2	21.4	20.1	43.1	42.3	31.6	57.1	18.1	7.6	56	3	51.2	46.7
	(20.7)	(24.3)	(33)	(20.4)	(25.9)	(24.7)	(36.3)	(37.5)	(28.8)	(36.1)	(24.9)	(11.5)	(39.1)	(5)	(30.8)	(38.7)
rhubarb	26.3	40.7	67.6	59.4	56.6	35.6	61.6	7.2	45.8	40	5.1	76.8	71.7	11.3	71.9	50
	(28.9)	(32.7)	(38.6)	(38.4)	(35.3)	(30.9)	(35)	(10.5)	(35.4)	(33.9)	(8.6)	(30.4)	(32.9)	(23.9)	(35.9)	(42.1)
parsley	55.3	34.7	55	53	21.1	25.5	28.6	9.5	13.4	15.3	6.8	8.9	18	5.8	74.9	63.2
	(30.8)	(31.7)	(43.6)	(42.3)	(23)	(31.8)	(35.8)	(19)	(17.6)	(24.8)	(10.4)	(11.4)	(21.5)	(12.3)	(28.8)	(38.3)

10.7 **Experiment 3: Multimodal SCs**

Table 90: SC-property-pairs with significant difference (p<.05) in mean property probability ratings between different typicality ratings for fruit SCs with multimodal typicality distributions with mean rating and number of participants (n) per mode and t-values.

subcategory	property	mode 1	mode 1 mean	mode 1 n	mode 2	mode 2 mean	mode 2 n	t12	mode 3	mode 3 mean	mode 3 n	t13	t23
	grew in the ground		11			25		2.4					
	will be boiled		35			20		2.6					
	has sugar		88			75		4.1					
plum	grew on a tree	1	96	17	3	86	7	3.9					
	will be eaten as dessert		76			65		2.4					
	tastes tart		49			37		2.3					
	will be used in cooking		51			34		2.9					
	will be eaten in soup		5			13		2.2					
fig	grew on a tree	1	95	9	4	78	9	3.1					
	will be eaten as dessert		84			72		2.7					
	will be made into juice		67			44		2.6					
	has sugar		74			48		3.2					
coconut	is sweet	4	77	8	7	35	8	7.1					
	grew on a tree		99			84		2.7					
	will be eaten as dessert		79			56		3.8					
	grew in the ground		7			7		0.1		21		3.2	2.5
	will be made into juice		81			70		1.7		65		2.6	0.8
	is sweet		80			56		3.2		33		4.7	2.3
prune	is juicy	1	73	7	3	42	7	5.6	4	37	5	6.4	1.0
	grew on a tree		84			55		2.5		67		1.3	0.8
	will be eaten as dessert		89			75		3.1		34		5.7	4.3
	tastes tart		43			39		0.5		77		4.3	3.8
	will be used in cooking		73			66		0.9		46		2.9	2.2
	grew in the ground		9			29		2.4					
	will be made into juice		56			86		3.9					
tomato	is sweet	3	56	7	7	36	7	2.5					
	is juicy		80			92		3.0					
	grew on a tree		15			4		2.5					
	has a heart		15			3		2.5					
	will be made into juice		62			30		3.3					
pumpkin	tastes tart	7	56	10	5	33	5	2.8					
	has a heart		8			2		2.7					
	grew on a plant		54			96		4.3					
	has roots		53			82		2.5					
	will be boiled		80			61		2.5					
	has sugar		86			51		4.2					
	is sweet		61			33		3.8					
rhubarb	is juicy	1	67	8	7	34	6	4.5					
	grew on a tree		5			0		3.2					
	will be eaten as dessert		87			53		4.2					
	will be used in cooking		95			66		4.0					
	grew on a plant		76			33		3.2					
	has roots		23			3		2.2					
avocado	will be made into juice	4	25	7	7	48	6	2.2					
	tastes tart		14			62		5.0					

Table 91: SC-property-pairs with significant difference (p<.05) in mean property probability ratings between different typicality ratings for vegetable SCs with multimodal typicality distributions with mean rating and number of participants (n) per mode and t-values.

subcategory	property	mode 1	mode 1 mean	mode 1 n	mode 2	mode 2 mean	mode 2 n	t12	mode 3	mode 3 mean	mode 3 n	t13	t23
	has roots		90			53		7.4					
	will be boiled		11			31		3.1					
	has sugar		55			27		2.5					
	has pips/seeds		7			18		2.3					
radish	is sweet	1	12	10	4	26	5	2.2					
	grew on a tree		3			33		5.5					
	will be eaten as dessert		4			13		7.2					
	tastes tart		16			47		3.4					
	will be used in cooking		43			79		3.3					
	will be eaten in soup		68			49		2.2					
corn	has roots	1	12	9	4	43	6	3.0					
	grew on a plant		86			50		3.2					
	will be eaten in soup		73			56		3.1					
	will be roasted		69			51		2.9					
onion	grew in the ground	1	95	13	3	84	8	2.4					
	will be boiled		54			32		3.7					
	has pips/seeds		3			7		2.4					
	tastes tart		37			16		2.9					
	will be roasted		66			84		2.3					
	is juicy		28			65		5.1					
eggplant	tastes tart	1	25	16	4	5	5	2.9					
	has a heart		14			1		2.9					
	will be used in cooking		85			97		2.6					
	grew in the ground		92			62		7.1					
green onion	grew on a tree	1	3	14	4	15	5	7.3					
	will be used in cooking		78			91		2.5					
mushroom	grew on a tree	1	5	9	5	28	5	4.1					
	grew on a plant		7			34		4.2					
	will be eaten in soup		10			15		2.0		29		4.2	2.1
	will be roasted		5			22		4.1		27		4.9	0.6
	has roots		4			21		7.1		8		1.7	3.0
pickle	will be made into juice	7	15	9	2	13	5	0.2	5	34	5	2.8	3.5
	grew on a tree		22			25		0.3		7		2.1	2.8
	will be eaten as dessert		3			8		2.7		19		4.1	1.7
	tastes tart		76			34		3.8		57		1.7	1.4
	will be eaten in soup		54			75		2.8					
	grew in the ground		90			74		2.6					
sweet potato	is juicy	1	21	11	4	7	7	2.9					
	tastes tart		10			30		3.4					
	has a heart		1			3		2.5					

Table 97 [cont.]: SC-property-pairs with significant difference (p<.05) in mean property probability ratings between different typicality ratings for vegetable SCs with multimodal typicality distributions with mean rating and number of participants (n) per mode and t-values.

subcategory	property	mode 1	mode 1 mean	mode 1 n	mode 2	mode 2 mean	mode 2 n	t12	mode 3	mode 3 mean	mode 3 n	t13	t23
garlic	will be eaten in soup	4	72	6	6	71	5	0.2	7	46	6	3.3	2.3
	has roots		62			62		0.0		30		2.6	2.2
	will be boiled		21			25		0.5		7		2.1	2.9
	will be made into juice		10			21		3.3		15		1.3	1.1
	has pips/seeds		7			4		2.7		5		1.8	0.8
	grew on a tree		6			5		0.5		1		3.7	2.8
	has a heart		8			3		3.0		2		4.2	0.9
	will be used in cooking		95			99		2.2		99		2.4	0.1
tomato	will be eaten in soup	7	80	10	1	55	9	4.1	4	48	5	4.2	0.7
	will be roasted		83			56		4.8		42		6.3	1.5
	will be made into juice		81			52		4.9		50		4.6	0.2
	is juicy		94			81		4.1		84		3.1	0.7
	grew on a tree		30			15		2.1		5		2.4	1.8
	will be eaten as dessert		8			23		3.6		7		0.2	2.2
	will be used in cooking		90			73		4.1		80		2.5	0.9
	grew on a plant		87			88		0.2		99		1.5	2.7
pumpkin	will be roasted	4	78	8	1	61	6	1.9	7	53	7	3.6	0.7
	has roots		62			37		1.8		33		2.5	0.3
	will be boiled		71			50		2.4		56		2.0	0.6
	has pips/seeds		97			93		0.9		78		2.8	1.8
	is sweet		73			33		6.2		57		2.5	2.4
	is juicy		51			38		1.2		59		0.8	2.2
	grew on a tree		5			2		2.1		8		1.2	2.8
	will be eaten as dessert		75			56		2.6		69		0.8	1.4
	tastes tart		32			18		1.6		42		1.2	3.0
	has a heart		3			12		2.2		7		1.6	0.9
	grew on a plant		52			81		2.6		60		0.7	1.8
rhubarb	will be eaten in soup	7	10	14	4	42	5	5.2					
	will be roasted		49			28		2.3					
	grew in the ground		73			29		4.3					
	will be made into juice		37			17		2.5					
	is juicy		57			38		2.2					
	will be used in cooking		84			59		2.7					
	grew on a plant		49			76		2.4					
avocado	grew in the ground	7	18	10	1	38	5	2.2					
	will be made into juice		16			57		4.6					
	is juicy		7			50		7.7					
	tastes tart		6			15		2.6					
	has a heart		2			34		4.7					

10.8 **Attribute-value-assignments**

Table 92: Attribute-value-assignments for the properties of the identified fruit and vegetable prototype from 5 researchers.

property	Attribute Decision	researcher 1	researcher 2	researcher 3	researcher 4	researcher 5
is eaten as dessert	CONSUMED-IN	consumed-in	usage		typically_used_in	ingredient in
is made into juice	CONSUMED-IN	consumed-in	usage		typically_used_in	ingredient in
is for soup	CONSUMED-IN	consumed-in	usage		typically_used_in	ingredient in
has a heart	HEART	internal-component	has a heart (yes, no), maybe as value of inner structure		internal_structure x heart	binary
grows on trees	HOW-GROWING	how-growing	origin (biological)		grow_location	where grows
grows on plants	HOW-GROWING	how-growing	origin (biological)		grow_location	where grows
grows in the ground	HOW-GROWING	how-growing	location while growing or origin (biological)		grow_location	where grows
has sugar	MAIN-NUTRITIONAL-COMPONENT	main-nutritional-value	main nutritional component / or contains sugar		nutritional_components x sugars	binary
is boiled	PREPARATION-METHOD	preparation	method of preparation	way of preparation	typical_cooking_methods	how prepared
is roasted	PREPARATION-METHOD	preparation	method of preparation	way of preparation	typical_cooking_methods	how prepared
has roots	ROOTS	external-structure	connection (of the accoring plant) into ground		plant plant roots	roots or stalks or neither
has pips/seeds	SEEDS	internal-component	seed (only for fruit)		seeds	means of reproduction
is sweet	TASTE	taste	taste / quality of taste	taste	taste	flavor
tastes tart	TASTE	taste	taste / quality of taste	taste	taste	flavor
is juicy	TEXTURE	texture	content of water		aridity	binary
is used in cooking	USED-IN-COOKING	use-in	usage or method of preparation	processing	typical_cooking_methods	binary

11 References

Adams, E. W. (1975). *The Logic of Conditionals*. *Synthese Library: Vol. 86*. Springer.

Ameel, E., & Storms, G. (2006). From prototypes to caricatures: Geometrical models for concept typicality. *Journal of Memory and Language*, *55*(3), 402–421. https://doi.org/10.1016/j.jml.2006.05.005

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, *13*(3), 263–308. https://doi.org/10.1016/0010-0277(83)90012-4

Balogh, K., & Osswald, R. (2021). A Frame-Based Analysis of Verbal Particles in Hungarian. In S. Löbner, T. Gamerschlag, T. Kalenscher, M. Schrenk, & H. Zeevat (Eds.), *Cognition and Mind: Vol. 7. Concepts, Frames and Cascades in Semantics, Cognition and Ontology* (pp. 219–237). Springer.

Barsalou, L. W. (1983). Ad Hoc Categories. *Memory & Cognition*(11 (3)), 211–227.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 629–654. https://doi.org/10.1037/0278-7393.11.1-4.629

Barsalou, L. W. (1987). The instability of graded structure. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorisation* (pp. 101–140).

Barsalou, L. W. (1992). Frames, Concepts, and Conceptual Fields. In Kittay, E., Lehrer, A. (Ed.), *Frames, Fields, and Contrasts* (pp. 21–74). Erlbaum.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, *7*(2), 84–91. https://doi.org/10.1016/S1364-6613(02)00029-3

Battig, W. F., & Montague, W. E. (1969). Category Norms for Verbal Items in 56 Categories: A Replication and Extension of the Connecticut Category Norms. *Journal of Experimental Psychology*(80 (3-2)), 1–46.

Berio, L., Latrouite, A., van Valin, R., & Vosgerau, G. (2017). Immediate and General Common Ground. In P. Brézillon, R. Turner, & C. Penco (Eds.), *Modeling and Using Context* (pp. 633–646). Springer International Publishing.

Berlin, B., & Kay, P. (1969). *Basic colour terms: their universality and evolution*. University of California Press.

Bhatia, R., & Davis, C. (2000). A Better Bound on the Variance. *The American Mathematical Monthly*, *107*(4), 353–357. https://doi.org/10.1080/00029890.2000.12005203

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a Brain-based Componential Semantic Representation. *Cognitive Neuropsychology*, *33*(3-4), 130–174. https://doi.org/10.1080/02643294.2016.1147426

Brooks, L. R., Norman, G. R., & Allen, S. W. (1991). Role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General*, *120*(3), 278–287. https://doi.org/10.1037//0096-3445.120.3.278

Brown, W. P., & Semrau, G. (1986). The Rated Typicality of Items in Semantic Categories: Some Irish Norms. *The Irish Journal of Psychology*, *7*(2), 123–146. https://doi.org/10.1080/03033910.1986.10557683

Chen, X. (2003). Object and Event Concepts: A Cognitive Mechanism of Incommensurability. *Philosophy of Science*, *70*(5), 962–974. https://doi.org/10.1086/377381

Cohen, B., & Murphy, G. L. (1984). Models of Concepts. *Cognitive Science*, *8*(1), 27–58. https://doi.org/10.1207/s15516709cog0801_2

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

Connolly, A. C., Fodor, J. A., Gleitman, L. R., & Gleitman, H. (2007). Why stereotypes don't even make good defaults. *Cognition*, *103*(1), 1–22. https://doi.org/10.1016/j.cognition.2006.02.005

Coolican, H. (2009). *Research Methods and Statistics in Psychology*. Routledge.

Corter, J., & Gluck, M. (1992). Explaining basic categories: Feature predictability and information. *Psychological Bulletin*, *111*(2), 291–303.

Cosmides, L., & Tooby, J. (1992). Cognitive Adaptations for Social Exchange. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (pp. 163–228). Oxford University Press.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2018). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*(3), 987–1006. https://doi.org/10.3758/s13428-018-1115-7

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, *40*(4), 1030–1048. https://doi.org/10.3758/BRM.40.4.1030

Devereux, B. J., Tyler, L. K., & Geertzen, J. (2014). The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior Research Methods*, *46*, 1119–1127.

Djalal, F., Storms, G., Ameel, E., & Heyman, T. (2017). What type of features do children associate with categories and how do they fare in predicting category judgements? *Acta Psychologica*, *178*, 114–123.

Douven, I. (2019). Putting prototypes in place. *Cognition*, *193*, 1–13.

Douven, I., & Gärdenfors, P. (2020). What are natural concepts? A design perspective. *Mind & Language*, *35*(3), 313–334. https://doi.org/10.1111/mila.12240

Efron, B., & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Estes, W. K. (1994). *Classification and cognition*. *Oxford psychology series: no. 22*. Oxford University Press; Clarendon Press.

Evans, Jonathan St. B. T., Handley, S. J., & Over, D. E. (2003). Conditionals and conditional probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(2), 321–335. https://doi.org/10.1037/0278-7393.29.2.321

Fodor, J. A. (1981). *Representations: Philosophical Essays on the Foundations of Cognitive Science*. The Harvester Press Limited.

Fodor, J. A., & Lepore, E. (1996). The red herring and the pet fish: why concepts still can't be prototypes. *Cognition*(58), 253–270.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. MIT Press.

Gati, I., & Tversky, A. (1984). Weighting Common and Distinctive Features in Perceptual and Conceptual Judgements. *Cognitive Psychology*, *16*, 341–370.

Gigerenzer, G. (2000). *Adaptive Thinking: Rationality in the Real World*. Oxford University Press.

Goldschmidt, A., Gamerschlag, T., Petersen, W., Gabrovska, E., & Geuder, W. (2017). Towards Verb Modification in Frames: A Case Study on German Schlagen (to hit). In H. H. Hansen, S. E. Murray, M. Sadrzadeh, & H. Zeevat (Eds.), *Logic, Language, and Computation* (pp. 18–36). Springer Berlin Heidelberg.

Goodman, N. (1972). *Problems and Projects*. The Bobbs-Merrill Company, Inc.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*(18 (4)), 441–461.

Hampton, J. A. (1982). A demonstration of intransitivity in natural categories. *Cognition*, *12*(2), 151–164.

Hampton, J. A. (1987). Inheritance of attributes in natural concept conjunctions. *Memory & Cognition*, *15*(1), 55–71. https://doi.org/10.3758/BF03197712

Hampton, J. A. (1988a). Disjunction of natural concepts. *Memory & Cognition*, *16*(6), 579–591. https://doi.org/10.3758/BF03197059

Hampton, J. A. (1988b). 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. (1993). Prototype Models of Concept Representation. In I. van Mechelen, J. A. Hampton, R. S. Michalski, & P. Theuns (Eds.), *Cognitive Science Series. Categories and concepts: Theoretical Views and Inductive Data Analysis* (pp. 67–95). Academic Press.

Hampton, J. A. (1995). Similarity-based categorisation: the development of prototype theory. *Psychologica Belgica*(35-2/3), 103–125.

Hampton, J. A. (1997a). Associative and similarity-based processes in categorisation decisions. *Memory & Cognition*, *25*(5), 625–640. https://doi.org/10.3758/BF03211304

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

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science*, *31*(3), 355–384. https://doi.org/10.1080/15326900701326402

Hampton, J. A. (2017). Compositionality and Concepts. In J. A. Hampton, Y. Winter, & 0002624 (Eds.), *Language, Cognition, and Mind: Vol. 3. Compositionality and Concepts in Linguistics and Psychology* (pp. 95–122). Springer International Publishing.

Hampton, J. A., & Gardiner, M. M. (1983). Measures of internal category structure: A correlational analysis of normative data. *British Journal of Psychology*(74), 491–516.

Hampton, J. A., & Jönsson, M. L. (2012). Typicality and Composition a Lity: the Logic of Combining Vague Concepts. In W. Hinzen, E. Machery, & M. Werning (Eds.), *The Oxford Handbook of Compositionality.* Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199541072.013.0018

Hampton, J. A., & Passanisi, A. (2016). When intensions do not map onto extensions: Individual differences in conceptualization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(4), 505–523. https://doi.org/10.1037/xlm0000198

Hartigan, J. A., & Hartigan, P. M. (1985). The Dip Test of Unimodality. *The Annals of Statistics*, *13*(1), 70–84.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2-3), 61-83; discussion 83-135. https://doi.org/10.1017/S0140525X0999152X

Hutchinson, J. W. (1983). On the locus of range effects in judgement and choice. *Advances in Consumer Research*, *10*, 305–308. https://www.acrwebsite.org/volumes/6130/volumes/v10/NA-10

Jönsson, M. L., & Hampton, J. A. (2012). The modifier effect in within-category induction: Default inheritance in complex noun phrases. *Language and Cognitive Processes*, *27*(1), 90–116. https://doi.org/10.1080/01690965.2010.544107

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809477.002

Kahneman, D., & Tversky, A. (1972). Subjective Probability: A Judgement of Representativeness. *Cognitive Psychology*(3), 430–454.

Kallmeyer, L., & Osswald, R. (2013). Syntax-driven semantic frame composition in Lexicalized Tree Adjoining Grammars. *Journal of Language Modelling*(1 (2)), 267–330.

Kittur, A., Holyoak, K. J., & Hummel, J. E. (2006). Ideals Aren't Always Typical: Dissociating Goodness-of-Exemplar From Typicality Judgements. *Abstracts of the Psychonomic Society 47th Annual Meeting*(11(120)).

Kornmesser, S., & Schurz, G. (2020). Analyzing Theories in the Frame Model. *Erkenntnis*, *85*(6), 1313–1346. https://doi.org/10.1007/s10670-018-0078-5

Larochelle, S., Richard, S., & Soulières, I. (2000). What some effects might not be: The time to verify membership in "well-defined" categories. *Quarterly Journal of Experimental Psychology: Section a*, *53*(4), 929–961.

Leslie, S.-J. (2008). Generics: Cognition and Acquisition. *The Philosophical Review*, *117*(1), 323–360. https://doi.org/10.1215/00318108-2007-001

Lesot, M.-J., Rifqi, M., & Benhadda, H. (2009). Similarity measures for binary and numerical data: a survey. *International Journal of Knowledge Engineering and Soft Data Paradigms*(1(1)), 63–84.

Löbner, S. (2014). Evidence for Frames from Human Language. In T. Gamerschlag, D. Gerland, R. Osswald, & W. Petersen (Eds.), *Studies in linguistics and philosophy: Vol. 94. Frames and concept types: Applications in Language and Philosophy* (Vol. 94). Springer.

Löbner, S. (2021). Cascades. Goldman's Level-Generation, Multilevel Categorisation of Action, and Multilevel Verb Semantics. In S. Löbner, T. Gamerschlag, T. Kalenscher, M. Schrenk, & H. Zeevat (Eds.), *Cognition and Mind: Vol. 7. Concepts, Frames and Cascades in Semantics, Cognition and Ontology* (pp. 263–310). Springer.

Malt, B. C., & Smith, E. E. (1982). The role of familiarity in determining typicality. *Memory & Cognition*, *10*(1), 69–75.

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior*, *23*(2), 250–269.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, *6*(4), 462–472. https://doi.org/10.3758/BF03197480

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559. https://doi.org/10.3758/BF03192726

Mervis, C. B., Catlin, J., & Rosch, E. (1976). Relationships among goodness-of-example, category norms, and word frequency. *Bulletin of the Psychonomic Society*, *7*(3), 283–284. https://doi.org/10.3758/BF03337190

Minsky, M. (1975). A Framework for Representing Knowledge. In P. H. Winston (Ed.), *The Psychology of Computer Vision* (pp. 211–277). McGraw-Hill.

Moreno-Martínez, F. J., Montoro, P. R., & Rodríguez-Rojo, I. C. (2014). Spanish norms for age of acquisition, concept familiarity, lexical frequency, manipulability, typicality,

and other variables for 820 words from 14 living/nonliving concepts. *Behavior Research Methods*, *46*(4), 1088–1097. https://doi.org/10.3758/s13428-013-0435-x

Murphy, G. L., & Medin, D. L. (1985). The Role of Theories in Conceptual Coherence. *Psychological Review*, *92*(3), 289–316. https://doi.org/10.1037/0033-295X.92.3.289

Nosofsky, R. M. (1984). Choice, Similarity, and the Context Theory of Classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*(10 (1)), 104–114.

Oberauer, K., & Wilhelm, O. (2003). The meaning(s) of conditionals: Conditional probabilities, mental models, and personal utilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 680–693. https://doi.org/10.1037/0278-7393.29.4.680

Ogden, C. K., & Richards, I. A. (1923). *The meaning of meaning*. Harcourt, Brace & World, Inc.

Osherson, D. N., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, *9*(1), 35–58. https://doi.org/10.1016/0010-0277(81)90013-5

Osswald, R., & van Valin, R. (2014). FrameNet, Frame Structure, and the Syntax-Semantics Interface. In T. Gamerschlag, D. Gerland, R. Osswald, & W. Petersen (Eds.), *Studies in linguistics and philosophy: Vol. 94. Frames and concept types: Applications in Language and Philosophy* (pp. 125–156). Springer.

Petersen, W. (2007). Representation of Concepts as Frames. In J. Skilters, F. Toccafondi, & G. Stemberger (Eds.), *The Baltic International Yearbook of Cognition, Logic and Communication: Vol. 2. Complex Cognition and Qualitative Science* (Vol. 2, pp. 151–170). https://doi.org/10.1016/B978-0-08-050913-6.50006-X

Pfeifer, N., & Kleiter, G. (2005). Coherence and nonmonotonicity in human reasoning. *Synthese*, *146*, 93–109.

Reichenbach, H. (1949). *The Theory of Probability: An inquiry into the logical and mathematical foundations of the calculus of probability*. University of California Press.

Rosch, E. (1973a). Natural Categories. *Cognitive Psychology*(4), 328–350.

Rosch, E. (1973b). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive Development and the Acquisition of Language* (pp. 111–144). Academic Press.

Rosch, E. (1975a). Cognitive Reference Points. *Cognitive Psychology*(7), 532–547.

Rosch, E. (1975b). Cognitive representations of semantic categories. *Journal of Experimental Psychology*(104 (3)), 192–233.

Rosch, E. (1978). Principles of Categorisation. In E. Rosch & B. Lloyd (Eds.), *Cognition and Categorisation.* https://doi.org/10.1016/S0083-6729(08)60516-6

Rosch, E. (2011). "Slow Lettuce": Categories, Concepts, Fuzzy Sets, and Logical Deduction. In R. Belohlavek & G. J. Klir (Eds.), *Concepts and Fuzzy Logic* (pp. 89–120). MIT Press.

Rosch, E., & Mervis, C. B. (1975). Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology*(7), 573–605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. https://doi.org/10.1016/0010-0285(76)90013-X

Rosch Heider, E. (1971). "Focal" color areas and the development of color names. *Developmental Psychology*, *4*(3), 447–455. https://doi.org/10.1037/h0030955

Rosch Heider, E. (1972). Universals in color naming and memory. *Journal of Experimental Psychology*, *93*(1), 10–20. https://doi.org/10.1037/h0032606

Ruts, W., De Deyne, S., Ameel, E., Vanpaemel, W., Verbeemen, T., & Storms, G. (2004). Dutch norm data for 13 semantic categories and 338 exemplars. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 506–515. https://doi.org/10.3758/BF03195597

Schröder, A., Gemballa, T., Ruppin, S., & Wartenburger, I. (2012). German norms for semantic typicality, age of acquisition, and concept familiarity. *Behavior Research Methods*, *44*(2), 380–394. https://doi.org/10.3758/s13428-011-0164-y

Schurz, G. (2001). What Is 'Normal'? An Evolution-Theoretic Foundation for Normic Laws and Their Relation to Statistical Normality. *Philosophy of Science*(68 (4)), 476–497.

Schurz, G. (2005). Non-monotonic reasoning from an evolution-theoretic perspective: Ontic, Logical and Cognitive Foundations. *Synthese*(146 (1/2)), 37–51.

Schurz, G. (2007). Human Conditional Reasoning Explained by Non-Monotonicity and Probability: An Evolutionary Account. In S. Vosniadou (Ed.), *Proceedings ofEuroCogSci07, The European Cognitive Science Conference 2007,* (pp. 628–633). Erlbaum. (Original work published 2007)

Schurz, G. (2011). *Evolution in Natur und Kultur: Eine Einführung in die verallgemeinerte Evolutionstheorie [Evolution in nature and culture: an introduction to generalised theory of evolution]*. Spektrum Akademischer Verlag.

Schurz, G. (2012). Prototypes and their Composition from an Evolutionary Point of View. In W. Hinzen, E. Machery, & M. Werning (Eds.), *The Oxford Handbook of Compositionality* (pp. 530–554). Oxford University Press. (Original work published 2012)

Schurz, G. (2013). *Philosophy of Science: A Unified Approach*. Routledge.

Schurz, G. (2015). *Wahrscheinlichkeit. Grundthemen Philosophie*. Walter de Gruyter.

Schurz, G. (2021). Evolution in Nature and Culture: Prospects and Problems of the Generalized Evolution Theory. *American Philosophical Quarterly*, *58*(1), 95–110. https://doi.org/10.2307/48600688

Schurz, G., & Votsis, I. (2014). Reconstructing Scientific Theory Change by Means of Frames. In T. Gamerschlag, D. Gerland, R. Osswald, & W. Petersen (Eds.), *Studies in linguistics and philosophy: Vol. 94. Frames and concept types: Applications in Language and Philosophy* (Vol. 94, pp. 93–109). Springer.

Schuster, A., Strößner, C., Sutton, P. R., & Zeevat, H. (2020). Stochastic Frames. In C. Howes, S. Chatzikyriakidis, A. Ek, & V. Somashekarappa (Eds.), *Proceedings of the Probability and Meaning Conference (PaM 2020).* The Association for Computational Linguistics.

Schwanenflugel, P. J., & Rey, M. (1986). The relationship between category typicality and concept familiarity: Evidence from Spanish- and English-speaking monolinguals. *Memory & Cognition*, *14*(2), 150–163.

Shepard, R. N. (1987). Toward a Universal Law of Generalization for Psychological Science. *Science*, *237*(4820), 1317–1323.

Smith, E. E., & Osherson, D. N. (1984). Conceptual Combination with Prototype Concepts *Cognitive Science*, *8*(4), 337–361. https://doi.org/10.1207/s15516709cog0804_2

Smith, E. E., Osherson, D. N., Rips, L. J., & Keane, M. (1988). Combining Prototypes: A Selective Modification Model. *Cognitive Science*(12), 485–527.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Storms, G. (2001). Flemish Category Norms for Exemplars of 39 Categories: A replication of the Battig and Montague (1969) category norms. *Psychologica Belgica*, *41*(3), 145–168.

Strößner, C. (2020a). Compositionality Meets Belief Revision: a Bayesian Model of Modification. *Review of Philosophy and Psychology*, *11*(4), 859–880. https://doi.org/10.1007/s13164-020-00476-8

Strößner, C. (2020b). Integrating conceptual spaces in frames. *IfCoLog Journal of Applied Logics*, *7*(5). http://www.collegepublications.co.uk/downloads/ifcolog00041.pdf

Strößner, C., & Schurz, G. (2020). The Role of Reasoning and Pragmatics in the Modifier Effect. *Cognitive Science*, *44*(2), e12815. https://doi.org/10.1111/cogs.12815

Strößner, C., Schuster, A., & Schurz, G. (2021). Modification and Default Inheritance. In S. Löbner, T. Gamerschlag, T. Kalenscher, M. Schrenk, & H. Zeevat (Eds.), *Cognition and Mind: Vol. 7. Concepts, Frames and Cascades in Semantics, Cognition and Ontology* (pp. 311–327). Springer.

Sutton, P. R. *Nominal Prototypes and Compositionality*.

Sutton, P. R. (2017). *Prototypes as Bayesian networks.* Workshop on probabilistic approaches to (prototype) concepts, Salzburg.

Sutton, P. R., & Filip, H. (2017). Individuation, reliability, and the mass/count distinction. *Journal of Language Modelling*, *5*(2), 303–356.

Tversky, A. (1977). Features of Similarity. *Psychological Review*(84 (4)), 327–352.

Tversky, A., & Gati, I. (1982). Similarity, Separability, and the Triangle Inequality. *Psychological Review*, *89*(2), 123–154.

Tversky, A., & Kahneman, D. (1982). Judgement under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 4–20). Cambridge University Press.

Urmson, J. O. (1970). Polymorphous concepts. In O. P. Wood & G. Pitcher (Eds.), *Modern studies of philosophy. Ryle* (pp. 249–266). Palgrave.

Uyeda, K., & Mandler, G. (1980). Prototypicality norms for 28 semantic categories. *Behavior Research Methods*, *12*(6), 587–595.

Votsis, I., & Schurz, G. (2012). A frame-theoretic analysis of two rival conceptions of heat. *Studies in History and Philosophy of Science Part a*, *43*(1), 105–114. https://doi.org/10.1016/j.shpsa.2011.10.010

Weiskopf, D. A. (2009). The plurality of concepts. *Synthese*, *169*(1), 145–173. https://doi.org/10.1007/s11229-008-9340-8

Wittgenstein, L. (1953). *Philosophical Investigations. Philosophische Untersuchungen*. Macmillan.

Wu, L.-l., & Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. *Acta Psychologica*, *132*(2), 173–189. https://doi.org/10.1016/j.actpsy.2009.02.002