Conceptual spaces and the strength of similarity-based arguments

Igor Douven, Shira Elqayam, Peter Gärdenfors, Patricia Mirabile

Conceptual Spaces and the Strength of Similarity-based Arguments*

Igor Douven
IHPST / CNRS / Panthéon–Sorbonne University

Shira Elqayam
School of Applied Social Sciences, De Montfort University

Peter Gärdenfors
Cognitive Science, Lund University

Patricia Mirabile
ILLC, University of Amsterdam

Abstract

Central to the conceptual spaces framework is the thought that concepts can be studied mathematically, by geometrical and topological means. Various applications of the framework have already been subjected to empirical testing, mostly with excellent results, demonstrating the framework's usefulness. So far untested is the suggestion that conceptual spaces may help explain certain inferences people are willing to make. The experiment reported in this paper focused on similarity-based arguments, testing the hypothesis that the strength of such arguments can be predicted from the structure of the conceptual space in which the items being reasoned about are represented. A secondary aim of the experiment concerned a recent inferentialist semantics for indicative conditionals, according to which the truth of a conditional requires the presence of a sufficiently strong inferential connection between its antecedent and consequent. To the extent that the strength of similarity-based inferences can be predicted from the geometry and topology of the relevant conceptual space, such spaces should help predict truth ratings of conditionals embodying a similarity-based inferential link. The results supported both hypotheses.

Keywords: argument strength; concepts; conceptual spaces; conditionals; inference; Inferentialism; similarity.

*The data and code for this paper are available at https://osf.io/8w5xa/?view_only=8567407c2e5f4fdca0772c9cb2e2f4ef.

1 Introduction

Logic studies the validity of arguments on the basis of their *form*. For instance, logic tells us that

> John is a sailor.
> All sailors can swim.
> John can swim.

is a valid argument because it is an instance of

$$\frac{Pa \qquad \forall x : Px \supset Qx}{Qa}$$

and the validity of this schema can be proven by simple set-theoretic means. That, in traditional logic, the only thing that matters is the syntactic form of the argument means, in particular, that the semantic contents of the variables are not considered. But while the importance of logic is hard to overstate, there are arguments that we deem valid, yet whose seeming validity cannot be explained by reference to logical form alone. For instance, an argument concluding that a marble is colored based on the premise that the marble is red appears perfectly all right. Yet this argument appears valid in virtue of its semantic content, rather than in virtue of its form: in judging it valid, we are exploiting our knowledge of the concepts RED and COLORED. In analytic philosophy, this problem has been handled by adding "meaning postulates" to the derivations (Carnap, 1952). Note, however, that in doing so we are simply explicating the conceptual knowledge we are relying on in judging the argument valid in the first place.

In everyday reasoning there are many examples of inferences that we consider more or less valid, even if they do not fall under the schemata of formal logic. For example, consider this argument:

> African elephants are highly social animals.
> Asian elephants are highly social animals.

Although not logically valid, this argument still seems to embody an inference one might reasonably make. Here, our judgment that the inference is reasonable crucially involves the notion of similarity: the concept of an African elephant and that of an Asian elephant are so similar that we have a prima facie inclination to deem properties holding of African elephants to hold of Asian elephants as well, and vice versa.[1]

Importantly, a reason to believe that one of the categories involved has a given property need not provide *equally strong* reason to believe that the other category has the property. It will matter how similar the categories are. For instance, we are much less inclined to infer that rhinos are highly social animals, let alone that manatees are, from the premise that African elephants are highly social animals. That is because rhinos are only somewhat similar, and manatees rather dissimilar, to African elephants.

[1] The qualification "prima facie" is important, because we realize that not *everything* true of the one category will be true of the other. For instance, African elephants tend to live in Africa, and we are clearly *not* willing to infer from this that Asian elephants live in Africa. Nevertheless, for many properties it will be the case that a reason to believe that one of the categories has it gives reason to believe that the other category has it, too.

What these examples indicate is that the semantic structure of concepts plays an important role in how we judge the validity of inferences. It is these semantic contents that generate our judgments of similarity. Hence, in order to study such inferences, we need a theory of how to describe semantic structure. In this article, we will base our analysis on the theory of conceptual spaces (Shepard, 1964, 1987; Nosofsky, 1986, 1987, 1989; Gärdenfors, 2000, 2014). In doing so, we focus on a particular version of that theory, to wit, the version proposed in Gärdenfors (2000). Our preference for this version over, for instance, Nosofsky's so-called Generalized Context Model is based on the outcomes of previous research (Douven, 2016; Douven et al., 2017) as well as on the fact that while there is a known proposal for modeling the kind of inference we are interested in using Gärdenfors' version (Osta-Vélez & Gärdenfors, 2020), it is not immediately evident what the analog of the proposal would be for Nosofsky's model; see Section 2.1 for details.[2]

Arguments of the above kind—in which we infer that items belonging to one category have a certain property from the premise that items belonging to another category have that property—have been studied extensively in cognitive psychology under the heading of "category-based induction." Empirical research in this area has consistently shown that the perceived strength of the inferential connection between premise and conclusion depends (among other parameters) on how similar people judge the involved categories to be, with strength of inferential connection typically increasing with greater similarity.[3]

Category boundaries also play a major role in slippery slope arguments, such as "If we accept voluntary ID cards in the UK, we will end up with compulsory ID cards in the future." Corner, Hahn and Oaskford (2011) developed a Bayesian account of slippery slope arguments, based on the (negative) utility of the consequent and a similarity-based probability account. They argued that such arguments were based on the categorical similarity between the end concepts (in this case, voluntary ID cards vs. compulsory ID cards), with an underlying process based on the implicit assumption that identifying the antecedent concept under a category increases the probability of the consequent concept being identified under the same category. They showed that manipulating similarity affected the perceived strength of slippery slope arguments, and that confidence in the categorization, moderated by similarity, predicted perceived argument strength.

Our own study will focus on a type of similarity-based argument that we introduce by dint of an example from Paris and Vencovská (2017):

> My son likes the movie *Toy Story*.
> ———————————
> My son likes the movie *The Sound of Music*.

[2] Research reported in Douven (2016) and Douven et al. (2017) also favors Gärdenfors' conceptual spaces framework over Hampton's (1998, 2007) Threshold Model, which is a different type of spatial model, and which otherwise also appears a promising starting point for formalizing the kind of similarity-based reasoning that is the topic of this paper. We should further mention Vigo and Allen (2009), we propose a non-spatial approach to connecting logical inference and similarity. These authors argue that, while in humans reasoning is strongly tied to language use, it is not inherently linguistic. They do so by showing how the logical operators can be cashed out in terms of subsymbolic processes computing similarity. Their notion of similarity appears closer to the one formalized by Tversky's (1977) set-theoretic approach than to the notion of similarity central to the conceptual spaces framework, to be detailed below. Also, Vigo and Allen are concerned with understanding deductive reasoning in terms of similarity while we are concerned with understanding a form of non-deductive reasoning in terms of similarity.

[3] This is so in single-premise category-based arguments. In multi-premise category-based arguments, in which we do not reason from one category to another, but from a number of categories to a further category, or to an overarching category (e.g., from robins, penguins, and ostriches to sparrows, respectively, to birds), the conclusion is typically better supported the more dissimilar the categories referred to in the premises are; see Osherson et al. (1990). There are limits, however, to how dissimilar the premise categories can be, as Sloman (1993) shows.

This argument embodies an inference one might or might not make when considering, for instance, whether the son would enjoy visiting the cinema to watch *The Sound of Music*. The argument involves the concept of being liked by the son, and it invites us to infer that a given object (*The Sound of Music*) falls into this concept from the premise that another given object (*Toy Story*) falls into the same concept. Here, too, similarity plays a key role. Whether we are willing to make the inference will depend on how similar, in our judgment, the two designated movies are. Carnap (1980) viewed arguments of this type as embodying a particular type of analogical reasoning, which he called "proximity-influenced."

There may be a concern that if similarity and concepts are so centrally involved in certain types of arguments, then the study of those arguments cannot attain the same high level of formal precision that is the hallmark of the logic literature, and which—it is generally believed—was achieved precisely by abstracting away from content and focusing strictly on syntax. Similarity, after all, would appear a vague and subjective notion, unlike logical form (Goodman, 1972; see Decock & Douven, 2011, for discussion of Goodman's arguments).

That is not necessarily so, however. The past decades have seen the development of a theory of conceptual spaces, mentioned above, which offers a framework in which similarity and concepts can be studied mathematically, by geometric and topological means. More specifically, in this framework concepts are represented as regions in similarity spaces, where the latter are constructed by statistical dimension-reduction techniques from similarity judgments, or confusion probabilities, or correlation coefficients, or similar data.

This approach to similarity and concepts has been subjected to experimental testing in the context of a variety of issues.[4] Very recently, Osta-Vélez and Gärdenfors (2020) have proposed to also use the framework for modeling category-based inductions. The idea of using conceptual spaces for modeling similarity-based arguments was in fact foreshadowed in work by van Fraassen (1967), Stalnaker (1979), and Carnap (1980), who at the time, however, did not have the same clear conception of conceptual spaces that we have today.

Osta-Vélez and Gärdenfors advance a hypothesis about how to derive the strength of a similarity-based argument from the geometry and topology of the relevant conceptual space. Broadly, their idea is that the strength of such an argument is a matter of premise–conclusion similarity and of premise and conclusion typicality (in senses to be made precise further on), where similarity and typicality can be determined from knowledge of the structure of the concepts in the relevant conceptual space. As said, their proposal concerns category-based induction, but we want to take it as a starting point for our research into proximity-influenced arguments, hypothesizing that Osta-Vélez and Gärdenfors' central idea applies to those arguments as well. More exactly, the work presented in the following was primarily motivated by the idea that, for instance, the strength of the argument from Paris and Vencovská's example depends on how similar the movies mentioned are judged to be, and on how typical they are for movies the son likes. Our main aim is to test this proposal empirically.

A secondary aim is related to the recent semantics of indicative conditionals—called "Inferentialism"—according to which the default interpretation of a conditional postulates a "strong enough" inferential connection between the conditional's antecedent and its consequent. As above, we investigate this connection using the structure of the concepts involved in the conditionals, again using similarity as a predictor. Inferentialism has recently received empirical support from a variety of experiments (Douven et al., 2018, 2020; Mirabile & Douven, 2020). However, in all experiments concerned with this semantics, the predictor or predictors were

[4] Useful summaries of empirical support so far reported for it are to be found in Gärdenfors (2014), Douven (2016a, 2019a, 2021a, 2021b), and Douven et al. (2017).

subjective measures (e.g., people's *judgments* of the strength of inferential connections). To the extent that the strength of proximity-influenced inferences can be predicted from the geometry and topology of a conceptual space—as per the above proposal—we should be able to predict truth ratings of conditionals embodying a proximity-influenced inferential link on the basis of precisely those (entirely objective) mathematical properties of the appropriate space. Whether that is so was another research question of the study presented in the following.

2 Theoretical background

The design of our study is informed by two theories: the theory of conceptual spaces and Inferentialism. We start by summarizing relevant work on conceptual spaces and describe Inferentialism in more detail than was done in the introduction.

2.1 The conceptual spaces framework

Central to the conceptual spaces framework is the thought that concepts can be represented in similarity spaces. Similarity spaces are mathematical objects, specifically one- or multidimensional metric spaces whose dimension or dimensions represent fundamental qualities along which items may be compared. Distances in such a space represent dissimilarities, in that the further apart items are (as represented in the space), the more dissimilar they are in the specific aspect the space is supposed to model (which could be color, taste, smell, and so on). In principle, there is a vast range of metrics that could be associated with a space. However, only two metrics have currency in the psychological literature, to wit, the city-block (or Manhattan) metric and the Euclidean metric. Given an n-dimensional space S, these are the instances of the following schema with $p = 1$ and $p = 2$, respectively:

$$\delta_S(x, y) \quad = \quad \sqrt[p]{\left(\sum_{i-1}^{n} |x_i - y_i|^p\right)},$$

with $x = \langle x_1, \ldots, x_n \rangle$ and $y = \langle y_1, \ldots, y_n \rangle$. Note that items very similar in one aspect could still be very dissimilar in other aspects; for example, items close to each other in taste dimensions might be far apart in color dimensions (e.g., a green apple and a red apple might taste very similar).

Similarity spaces are commonly constructed from similarity judgments, or from so-called confusion probabilities (which are data about how likely two distinct stimuli are mistaken to be identical when flashed consecutively to participants), or from correlation coefficients (as in Douven, 2021a, for instance). In a pre-processing step, such data are transformed into distances. Then a multidimensional scaling (MDS) procedure, or a variant thereof (like non-negative matrix factorization or principal component analysis), is applied to those data to generate a metrical space. In this kind of procedure, the goal always is to obtain a space that is low-dimensional, adequately fits the data, and is interpretable, in that we can relate the space's dimensions to fundamental qualities that the stimuli used can be said to possess (Borg & Groenen, 2000; Hout, Papesh, & Goldinger, 2013; Abdi & Williams, 2010).

MDS and related procedures come without a guarantee of success: there may simply be no low-dimensional spatial representation that adequately fits our data, and if there is, we may struggle to come up with a meaningful interpretation of its dimensions. Nonetheless, by now cognitive psychologists have uncovered a number of similarity spaces that do satisfy all desiderata. More importantly for present purposes, they have been able to build structures on

top of some of those spaces that render them suitable for use as *conceptual* spaces, meaning that they not only represent similarity relations but also concepts, understood as the mental correlates of words.

Opinions vary somewhat as to how we are to obtain a conceptual space from a similarity space, but a leading idea uses a combination of prototype theory and the mathematical technique of Voronoi tessellations (Gärdenfors, 2000, 2014). According to the former, not all instances of a concept are equally representative of it, and the one that best represents it is its prototype (Rosch, 1973, 2011). A Voronoi tessellation of a given space divides that space into disjoint cells, where each cell is associated with exactly one so-called generator point and contains precisely those points in the space that are at least as close to that cell's generator point as they are to any other cell's generator point.[5] One obtains a conceptual space from a similarity space by locating in the latter the prototypes of various concepts and using these to generate a Voronoi tessellation of the space. The cells of that tessellation then form the concepts represented by the space.

In a recent extension of this framework, concepts may, instead of prototypes, also have *prototypical regions*. For instance, there would appear to be no unique shade of red that is *the* best representant of the concept RED, no unique shade that strikes us as being typically red. Early support for this idea is to be found in Berlin and Kay's (1969) seminal work on basic color categories. It is further supported by findings reported in Douven (2016a, 2019) and Douven et al. (2017). The thus extended framework has been used to explain the vagueness of some of our concepts (Douven et al., 2013) and to complete Kamp and Partee's (1995) proposal for defining notions of graded membership and partial truth (Decock & Douven, 2013; Douven & Decock, 2017, Verheyen & Égré, 2018).

Douven (2016a) reports the results of a number of studies meant to test the accounts of vagueness and graded membership. The study to be presented in the following builds on the first two studies from that paper. More specifically, we use the similarity space for container-like shapes that came out of the first study and the data gathered in the second study concerning prototypical regions in that space.

All studies reported in Douven (2016a) used the same 49 stimuli, which are shown in Figure 1. Over 1000 participants took part in the first study. In this study, each participant was shown 25 pairs consisting of two different stimuli, where the pairs were randomly chosen per participant. The participants were tasked to rate the similarity of the members of each pair on a 10-point Likert scale. The results from this study were aggregated across participants, and these aggregates formed the input for an MDS procedure. From this procedure, a three-dimensional space with a city-block metric defined on it emerged as scoring best on all relevant model-fit criteria, and also as scoring well, absolutely speaking (see Douven, 2016a, for details).

The second study from Douven (2016a) had the purpose of locating the prototypical regions for concepts that could plausibly be interpreted in the shape space, notably the concepts of bowl, cup, mug, pot, and vase. There was not much support for the claim that any of the shapes in Figure 1 is typical for cups, mugs, and pots. By contrast, most participants deemed various shapes typical for vases and various other shapes typical for bowls. Figure 2 shows the locations of these shapes in the three-dimensional city-block space, together with their convex hulls, where the typical vases span the purple hull and the typical bowls the green hull.

To be sure, there exist several other models of categorization that build on spatial representations, beginning with Shepard's (1964) model. Among later developments one finds Nosofsky's (1986, 1987, 1989; also Nosofsky & Zaki, 2002) Generalized Context Model (GCM)

[5] For formal details, see Okabe et al. (2000).

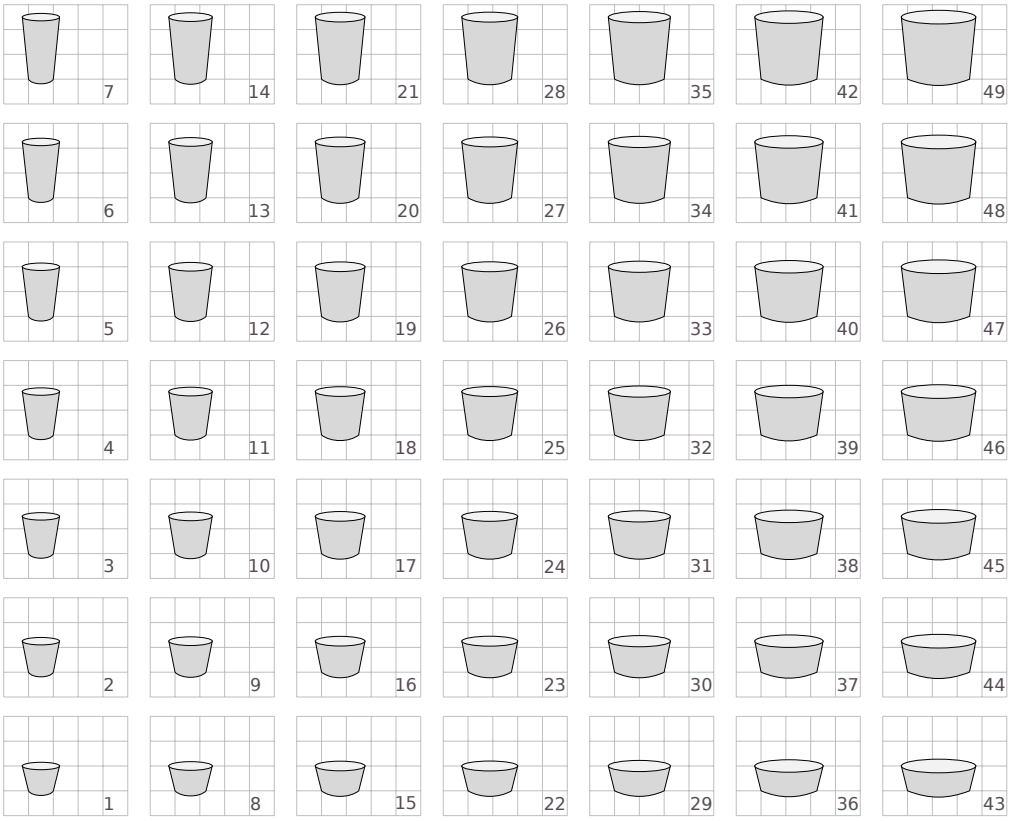

Figure 1: The 49 figures that were used in all studies reported in Douven (2016a). (The numbers in the bottom-right square of each grid did not appear in the pictures used in those studies, but merely served as labels for the shapes in Douven, 2016a.)

that was already briefly mentioned in the previous section. This model assumes similarity to be measured by an exponential decay function, while our model is not dependent on this kind of assumption. More importantly, his model builds on exemplars rather than on prototypes and aims to predict categorization probabilities rather than typicality. By contrast, and as said, the model we are testing builds on a prototype structure of categories. Neither Nosofsky's model nor Shepard's is applicable in the current context since they do not give any predictions concerning the implications we are testing. This is not to say that it is impossible to develop accounts of analogical reasoning using those models. Indeed, such an account has been developed in a different context: in psycholinguistics, the GCM has been used to model the analogical route to inferring inflectional morphology (e.g., Hahn & Nakisa, 2000). Speakers are typically able to generalize from a known regularity in morphology to a novel stimulus. For example, the past tense of "ring" is "rung"; it fits with a category of irregular English verbs such as "swing"–"swung," "string"–"strung," and so on. (The term "minor rule" is sometimes used to refer to such "regular irregularities.") When presented with a non-word such as "spling," speakers typically draw on implicit knowledge of such minor regularities to produce its past form as "splung" (rather than add the regular -ed suffix to produce "splinged"). Such productivity is graded: the more phonologically similar the non-word is to the prototypical pattern, the more likely that speaker will produce this pattern. Whether such an analogical route is sufficient as a single route model is a matter of contention in the literature, with some authors favoring

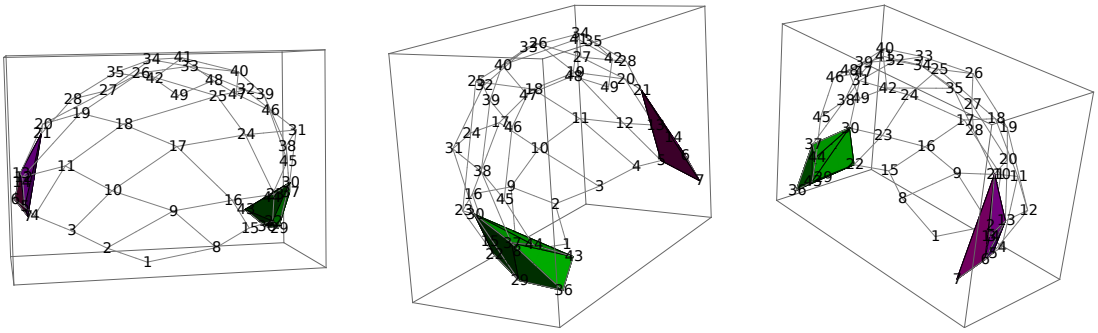

Figure 2: This is Figure 6 from Douven (2016a), showing the three-dimensional city-block space from different viewpoints. The numbers correspond to the labels used in Figure 1. In purple, the convex hull of majority choices of typical vase shapes (encompassing shapes), and in green, that of majority choices of typical bowl shapes (encompassing shapes).

instead a dual-route model, in which analogy only covers irregular forms, while regular forms are governed by rules (see also Albright & Hayes, 2003; Dawdy-Hesterberg & Pierrehumbert, 2014). However, no such model has been developed for anything resembling our current work.

It is also to be noted that both Douven (2016a) and Douven et al. (2017) compared the GCM with the conceptual spaces approach. Both papers were concerned with graded membership and obtained highly accurate predictions for the degrees to which the shapes in Figure 2 were deemed to represent a vase or deemed to represent a bowl and, respectively, the degrees to which various color shades in the blue/green border region in color space were deemed to be blue or deemed to be green. Both aforementioned papers highlighted an important advantage of the conceptual spaces approach over the GCM, to wit, that, on the former model, predictions of degrees of category membership *follow directly* from the geometry and topology of the relevant space or spaces, *without any parameter estimation being required*, while the GCM requires one to estimate the values of a considerable number of parameters, including the typicality gradient (which determines the steepness of the exponential decay function), a response bias parameter for each category involved, an attention weight for each relevant dimension of evaluation, and a response scaling parameter. That alone makes the conceptual spaces approach predictively far more specific than the GCM: rather than asserting that values can be found for certain parameters which will yield degrees of membership that match the observations, the conceptual spaces approach allows one to directly predict these degrees of membership. Consequently, even if the predictions obtained from the GCM are as accurate as those obtained from the conceptual spaces approach—which, as said, were *highly* accurate—then considerations of simplicity, as formally incorporated, for instance, in AIC, BIC, and similar model selection criteria, still favor the conceptual spaces approach.

Furthermore, Bellmund et al. (2018) show that neural mechanisms in the hippocampal system which are exploited in spatial navigation generalize across information domains, thereby supporting a wide spectrum of cognitive functions including concept formation and memory. The place and grid cell population codes represent variable dimensions of cognitive spaces. This mapping system enables a multitude of stable cognitive spaces at different resolutions and hierarchical levels. The spatial representations of the hippocampal formation thereby support flexible cognition and behavior. This provides additional strong empirical evidence for the validity of conceptual spaces as modelling tools.

2.2 Inferentialism

The present paper has a secondary aim related to conditionals. The connection of the foregoing to conditionals is, in fact, straightforward from the perspective of a recent account of conditionals that puts the inferential connection between antecedent and consequent center stage. According to this view, whether we evaluate as true the conditional, "If my son likes *Toy Story*, he'll like *The Sound of Music*," depends on the perceived strength of the argument from Paris and Vencovská's (2017) example stated in the introduction. The account predicts that the more compelling this argument appears to us, the likelier we are to deem the corresponding conditional true.

According to the account meant here—Inferentialism—the truth of a conditional requires the existence of a compelling, "strong enough" argument (in the sense of Simon, 1982) from the conditional's antecedent (plus background knowledge) to its consequent, where the antecedent is pivotal in the argument, meaning that, without the antecedent, the argument for the consequent is no longer compelling (Krzyżanowska, Wenmackers, & Douven, 2014; Krzyżanowska, 2015; Douven, 2016a, 2017; Douven et al., 2018; see also Oaksford & Chater, 2010, 2013, 2014, 2017, 2020, Vidal & Baratgin, 2017, and van Rooij & Schulz, 2019, Krzyżanowska, Collins, & Hahn, 2021). Psychologically, this inference is driven by relevance and bounded by satisficing: that is, people represent by default the inferential connection as relevant, and they satisfice, rather than optimize (again in the sense of Simon) on the strength of the connection. We called the psychological theory based on Inferentialism "Hypothetical Inferential Theory" (HIT, for short). One implication of HIT is that reasoners are also susceptible to the same biases characteristic of inference more generally when evaluating conditionals' truth. Indeed, Douven et al. (2018) found that truth evaluation was strongly affected by the conditional's consequent, in much the same way that inference generally is affected by belief in the conclusion of an argument (Evans et al., 1983; Evans, 2006, 2007). In other words, there is a robust belief bias analogue in evaluations of the truth of conditionals.

The idea that the truth of a conditional requires an inferential connection between its component parts harks back to the Stoics (Kneale & Kneale, 1962), and is also found in the later writings of, among others, Mill (1843/1872), Ramsey (1929/1990), and Mackie (1973). That the idea has nonetheless never been widely popular is mainly because—critics have argued—there are conditionals that appear true without their consequent being inferable from their antecedent. For instance, it requires little effort to imagine a context in which we would regard the statement, "If Betty misses her bus, she will be late for the movies," as being true, yet in which we cannot rule out entirely that Betty is transported from her present location to the cinema after missing the bus but still before the beginning of the movie.

As Krzyżanowska, Wenmackers, and Douven (2014) point out, however, this criticism presupposes that we are to interpret "inference" as meaning *deductive* inference, an interpretation to which the idea that the truth of a conditional requires the existence of an inferential connection is not wedded. It may instead refer to a broader notion of inference, one that encompasses other forms of inference besides deduction, most notably induction and abduction, and indeed also proximity-influenced inference, which is our main current topic.[6]

[6]To forestall misunderstanding, we note that Inferentialism—as stated in Krzyżanowska, Wenmackers, and Douven (2014)—is limited to standard indicative conditionals, excluding so-called nonconditionals (Lycan, 2001) such as speech act conditionals ("If you're hungry, there are cookies on the table") and non-interference conditionals ("If hell freezes over, Alice will not marry Bob"), as well as subjunctive conditionals and concessives (i.e., "even if" conditionals, which are sometimes also expressed without "even"). As a result, criticisms that accuse Inferentialism of being unable to account for, for instance, concessives (e.g., Mellor & Bradley, 2021) are misguided. As a further

There is already considerable experimental support for the idea that inferential connections (broadly construed, so as to include non-deductive inferential connections) are key to how people evaluate conditionals. Skovgaard-Olsen et al. (2019) report the finding of a pattern of individual differences with respect to people's probability judgments of conditionals in which a majority conformed to this approach. Furthermore, the re-analysis in Douven et al. (2020) of the data from Douven et al. (2018) showed truth ratings of conditionals to be better explained by taking the strength of inferential connections into account than by any of the standard semantics of conditionals (such as the material conditional account and the possible worlds semantics from Stalnaker, 1975), which assign no role to such connections. And Mirabile and Douven (2020) found that endorsement rates for modus ponens and modus tollens were more accurately predicted by the strength of the inferential connection between the major premise's component parts than by the probability of that premise's consequent given its antecedent.[7]

3 Plan

To our knowledge, the first published suggestion that (what are now called) conceptual spaces can play a role in explaining the pre-theoretical validity of certain types of non-deductive inference is to be found in van Fraassen (1967). In considering why we are licensed to infer that an object is red from the premise that it is scarlet, van Fraassen argues that, in color space, the region representing scarlet is included in the region representing red. Apparently, van Fraassen was unaware of the work on similarity spaces that had then just begun in cognitive psychology (see, e.g., Shepard, 1964). At any rate, he was working with what we would now consider an inadequate conception of conceptual spaces; for instance, he simply identifies color space with the color spectrum.

We find a related idea in Carnap's work on inductive logic from the 1960s, work that was published posthumously, as Carnap (1980). For much of his career, Carnap had been trying to define inductive inference in strictly syntactic terms, a project he gave up only in his last writings on inductive logic. There, he introduces *attribute spaces*, which are basically just conceptual spaces, abstractly described, although Carnap's conception was closer to ours than van Fraassen's was; for instance, color space for him is a three-dimensional double cone, which indeed somewhat resembles our currently best color spaces (CIELab space and CIELuv space; see Fairchild, 2013). Carnap is concerned with explicating the notion of confirmation (rather than with non-deductive arguments) and, among other things, he proposes that the degree to which the finding that an object o has property P supports the hypothesis that another object o' has P as well depends on how similar o and o' are, which he takes to be given objectively by their distance in the relevant attribute space (Carnap, 1980, Section 17 C).

Although there was a vast increase in interest in conceptual spaces after 1970 (the year of Carnap's death), the idea of connecting such spaces to the topic of non-deductive inference was only very recently taken up again, by Osta-Vélez and Gärdenfors (2020). The broad idea underlying their proposal is that the strength of concept-based arguments depends on three

aside, we note that Douven (2016b, p. 38 f) already pointed out that one could easily extend Inferentialism to cover subjunctive conditionals. And an inferentialist account of concessives might define "[Even] if φ, ψ" to be true if, and only if, there is a compelling argument for ψ from background premises alone and also from those premises revised (in the sense of Alchourrón, Gärdenfors, & Makinson, 1985) with φ (i.e., given one's current background knowledge, there is a compelling argument from φ to ψ, but φ would be redundant in that argument).

[7] For further evidence, see Krzyżanowska, Collins, and Hahn (2017, 2021), Vidal and Baratgin (2017), Krzyżanowska and Douven (2018), Rostworowski, Pietrulewicz, and Będkowski (2021), and Stewart et al. (2021). But cf. Skovgaard-Olsen et al. (2017).

elements: premise–conclusion similarity, premise typicality, and conclusion typicality. They make this exact using the conceptual spaces framework, as follows:

$$\log \mathbb{E}\left[S(X \rightarrow Y)_Z\right] \;=\; \mathrm{sim}(X, Y) + a\ \mathrm{sim}\left(X, p^Z\right) + b\ \mathrm{sim}\left(Y, p^Z\right).$$

This says that the logarithm of the expectation that Y has S if X has S, where X and Y are concepts both falling within a more encompassing concept Z, is equal to the (weighted) sum of the similarity between X and Y—which is the inverse of the distance between the regions representing X and Y in the relevant conceptual space—and the Z-typicality of X and Y, that is, the distance between X, respectively, Y and the prototype of Z, p^Z. The coefficients a and b weight the contributions made by the Z-typicality of X and Y and are to be estimated from the data. The known data about category-based induction suggest that Osta-Vélez and Gärdenfors' proposal is along the right lines, albeit that some data suggest an influence of premise typicality but not conclusion typicality (e.g., Rips, 1975), while others suggest the opposite (Hampton & Cannon, 2003). One possible explanation for this is that the coefficients a and b can vary across conceptual spaces and thus require indexing.

As mentioned, in this paper we focus on an inference type closely related to category-based induction, to wit, what following Carnap we call "proximity-influenced arguments," that is, arguments of the type seen in Paris and Vencovská's (2017) example discussed in the introduction, in which it is inferred that the son will like *The Sound of Music* from the premise that he liked *Toy Story*. Taking our cue from Osta-Vélez and Gärdenfors' proposal, we hypothesize that people's judgments of the strength of such arguments is a function of premise–conclusion similarity, premise typicality, and conclusion typicality.

To underline the close kinship with category-based induction, note that if, following Gärdenfors (2000), we conceive of an object as a special type of concept, then a proximity-influenced argument can be thought of as a special kind of category-based argument. Instead of projecting a property from one category (or concept) to another, proximity-influenced arguments project a property from one object (special type of concept) to another. In the latter type of argument, too, similarity may be crucial in determining their strength. And even though, to our knowledge, no one ever proposed that, in the case of proximity-influenced arguments, the legitimacy of the projection might involve matters of typicality—for instance, whether *Toy Story* is the kind of movie the son *typically* enjoys watching (e.g., animation movies)—in keeping close to Osta-Vélez and Gärdenfors' proposal we conjecture that, *as a matter of fact*, issues of typicality *will* factor into people's willingness to engage in proximity-influenced reasoning, at least to some degree. In short, we hypothesize that whether people are willing to make a proximity-influenced inference depends on (i) how similar the objects designated in premise and conclusion are to each other; (ii) how typical for the projected property the former object is; and (iii) how typical for the projected property the latter object is.

The plan was to test this hypothesis using the shape space from Douven (2016a), which means that our materials will be about vessels rather than movies, and the property we will project in the argument will be that of being a vase rather than being liked by a person. But structurally the arguments our materials present to participants will be identical to Paris and Vencovská's example. As pointed out, with each such argument we can associate a conditional that has the argument's premise as its antecedent and the argument's conclusion as its consequent. In our materials, we matched arguments with corresponding conditionals, using the former in one task and the latter in another. These tasks allowed us to test both the aforementioned hypothesis about proximity-influenced arguments and Inferentialism. Specifically, we tested whether the perceived strength of a proximity-influenced argument

was a reliable predictor of the truth rating of the corresponding conditional. It is to be noted that this does not amount simply to a replication of previous work (e.g., Douven et al., 2018), albeit for proximity-influenced inference; we are now also able to test whether truth ratings for conditionals embodying a proximity-influenced inference can be predicted on the basis of the metric and topological properties of the relevant conceptual space (in our case, the said shape space).

Thus, our study was aimed at testing two broad hypotheses, to wit, first, that the strength of proximity-influenced arguments is a function of (i) the distance in the relevant conceptual space between the object designated in the premise and the object designated in the conclusion, as well as (ii) how typical these objects are of the projected property; and second, that truth ratings of conditionals embodying proximity-influenced inference can be predicted on the exact same basis. From these hypotheses, we derived several more specific predictions:

1. Based on Osta-Vélez and Gärdenfors' proposal, we predicted a main effect of similarity on inference strength.
2. Based on Osta-Vélez and Gärdenfors' proposal, we also predicted a main effect of premise typicality and a main effect of conclusion typicality, also on inference strength.
3. Based on HIT / Inferentialism, we expected inference strength to be a strong predictor of truth evaluation.
4. Furthermore, given 1 and 3, we predicted a main effect of similarity on truth evaluation.
5. As for the effect of typicality on truth evaluation, the prediction is not entirely straightforward. There are two possibilities:
   (a) Given 2 and 3, we should predict a main effect of both antecedent typicality and consequent typicality on truth evaluation
   (b) Given previous findings on belief bias, we can predict a main effect of antecedent typicality but not of consequent typicality.

It is worth pointing out that, first, prediction 3 is the linchpin connecting Osta-Vélez and Gärdenfors' proposal with Inferentialism and HIT, and second, our study's design can distinguish between prediction 5a and prediction 5b.

4 Study

4.1 Method

4.1.1 Participants

Participants were recruited via Prolific (https://www.prolific.co/). A total of 113 participants completed the study. Of these, 20 were excluded for having had advanced training in logic, or a dyslexia diagnosis, or for not indicating English as their native language, or for failing an attention check.[8] This left us with 93 participants (66 women, 24 men, 3 non-binary/unspec-

[8] There were four attention checks: (1) In the demographics section, participants were given a list of hobbies, and instructed to write "I read the instructions" in the "Other" box, a procedure we adapted from Pennycook et al. (2014). (2) After the argument strength part of the study, participants were shown a photograph of flamingos and asked to count how many there were. (3) After the truth rating task of the study, participants were presented with a drawing of three colored balls, and asked to indicate the color of the leftmost one. (4) At the end of the study, participants were asked if they had answered seriously, a procedure we adapted from Aust et al. (2013). Participants were excluded if they failed any of the checks.

ified gender individuals; M_{age} = 35.20, SD_{age} = 12.23). Participants spent on average 514.27 (± 200.70) seconds on the survey. They received £1.88 each for their participation.

4.1.2 Materials and procedure

The study was run online using the Qualtrics platform (https://www.qualtrics.co/). It consisted of two tasks, which were presented in a counterbalanced order: an argument strength task, and a truth rating task. Both tasks drew on the shapes shown in Figure 1. To minimize the risk of carry-over effects, the participants were asked to answer the demographic questions in-between the two tasks.

At the beginning of the study, the participants received the following instructions:

> Imagine the following scenario: This is the first day of your summer job in a pottery shop. Bowls and vases are displayed on separate shelves, and your new boss is very particular about displaying vessels on the correct shelves: vases with vases and bowls with bowls. A large consignment of vessels has just arrived, and the vessels are shown in the picture below. [Here, the participants were shown Figure 1, though without the numbers appearing in that figure.] The vessels need to be sorted to the appropriate shelves, but it isn't always obvious which is which. Your boss is off to lunch and has left the sorting to you. You don't have anyone to ask, but there are already some other bowls and vases sorted on the appropriate shelves so you can take your hints from them.

Then began one of the two tasks, in each of which they were asked twelve questions. Each question concerned a pair of the 49 shapes shown in Figure 1. For each of the questions of the first task the participants received—whichever that was—a pair of different shapes was randomly drawn from the said collection. We ensured that in the second task the participant received, they would see precisely those 12 pairs of shapes that they had seen in the first task. Each question appeared on a separate page (i.e., screen), which showed the selected pair of shapes side by side.

In the argument strength task, participants were required to suppose that the vessel that appeared on the left was a vase and then to indicate whether that gave them reason to believe that the vessel on the right was a vase as well. The response had to be given on a 7-point Likert-scale, with only the anchors being labeled, as "Definitely does NOT give me reason to believe" and "Definitely does give me reason to believe." In the truth rating task, each question presented the participant with a conditional, "If the vessel on the left is a vase, so is the vessel on the right." They were then asked to indicate whether they thought this was true, false, or neither. They were instructed to choose the latter if they thought that, for whatever reason, it was impossible to evaluate the conditional as true or false.

4.2 Results and discussion

The data used in the analysis consisted of the responses from the participants in our study as well as of some data taken from Douven (2016a).[9] As for the former, in each of the two tasks from the study, the 93 participants answered 12 questions. This means that, in principle, $93 \times 12 = 1116$ unique pairs of shapes, out of the $\binom{49}{2} = 1176$ possible pairs, could have been viewed by participants. It turned out that, in actuality, 890 unique pairs had been presented

[9] The data from Douven (2016a) are publicly available at https://www.sciencedirect.com/science/article/abs/pii/S0010027716300622.

to participants. Of these, 697 pairs had been presented to one participant, 163 to two, 27 to three, and three had been presented to four participants. For each pair of participant and pair of shapes seen by that participant, we had one truth rating of a conditional and one inference strength rating of the corresponding argument. As for the data from Douven (2016a), these consisted of (i) the coordinates of the 49 shapes in Figure 1 in the three-dimensional city-block space that had come out of the MDS procedure from the analysis of Study I reported in the said paper; and (ii) the coordinates of the typical vase shapes, as had been identified in Study II from the same paper. From these data, we calculated the distance between the shapes in each pair that had occurred in our study, and we calculated the distance to the nearest typical vase shape for each of the 49 shapes.

4.2.1 Predicting truth ratings on the basis of inference strength

Earlier work had found a strong connection between truth ratings of conditionals and the strength of the inferential connection between conditionals' component parts. We started with a test of prediction 3, the linchpin between the two theoretical approaches bridged in this study. We therefore first investigated whether our new data were consistent with those earlier findings, in particular, whether the participants' truth ratings could be predicted on the basis of their judgments of the strength of the arguments corresponding to the rated conditionals.

In the truth rating task, participants were given the option "Neither/nor" which they were instructed to choose if, for whatever reason, they could not say whether a conditional was, in their judgment, true or false. The "Neither/nor" category is not quite on a par with the "True" and "False" categories and is indeed rather unspecific: a "Neither/nor" response could indicate that a participant deems a conditional's truth value to be indeterminate, or to have a third truth value somehow in-between truth and falsity (as they can have in some three-valued logics), or simply to be undecided about which of "True" and "False" to choose. Therefore, we followed the procedure of Douven et al. (2018) and conducted separate analyses for the "True" versus "False" responses and the full set of responses.

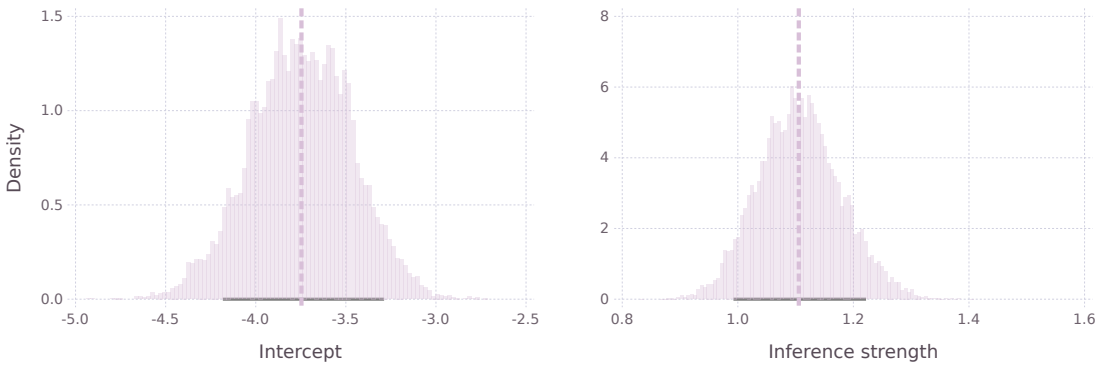

Figure 3: Posterior distributions for the intercept and inference strength (in log odds). The dotted vertical lines show the medians and the dark horizontal lines at the bottom show the 89 percent HDI.

In the first analysis, we ran a Bayesian mixed-effects logistic regression, using the `Turing.jl` package for the high-performance computing language Julia (Bezanson et al., 2017).[10,11] The

[10] For detailed explanations of the Bayesian models we are using in this paper, see Kruschke (2015) and McElreath (2020), which are excellent textbooks on Bayesian statistics.

[11] Here and elsewhere, we also ran separate analyses for the group that had received the argument strength task

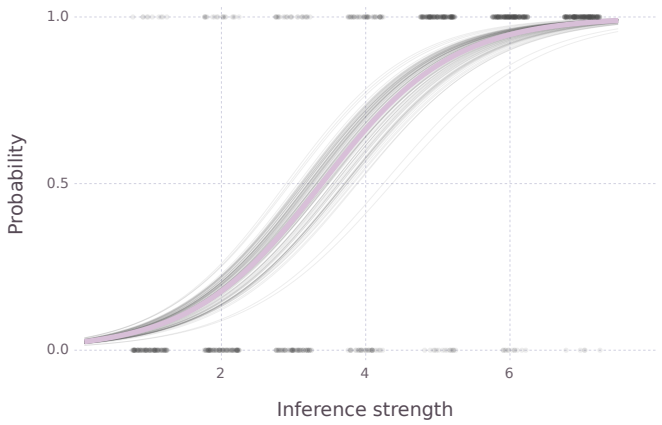

Figure 4: The colored line shows the predicted median probability of truth based on the fixed effect. The thin gray lines show the predictions of the individual random effects from all participants. The data are shown as gray dots, with jitter added to enhance visibility.

model had the true/false responses from the truth rating task as a binary response variable, the responses from the inferential-strength rating task as fixed effect, and per-participant random intercepts and random slopes. Informed by the results from Douven et al. (2018, 2020), we chose as priors standard normal distributions both for the fixed effect and for the fixed effect intercept, $\text{Cauchy}_+(0, 2.5)$ distributions for the standard deviations in the random effects, and an LKJ(2) distribution for the correlation matrix of the random effects. MCMC diagnostics gave no reason for concern, indicating sufficient mixing of the chains, sufficiently high bulk and tail effective sample size values, and an $\widehat{R}$ convergence diagnostic of 1.00 for all parameters.

The median of the posterior distribution for the intercept equaled −3.75 (89 % HDI [−4.18, −3.29]) and that of the posterior distribution for inference strength equaled 1.11 (89 % HDI [0.99, 1.22]).[12] See Figure 3 for plots of the full posterior distributions. Figure 4 shows the predicted probabilities together with the data.

To interpret these findings, consider that $\exp(-3.75) \approx 0.024$, which is the odds for a conditional to be judged true if the strength of the inferential connection between its component parts is at the (hypothetical) value of 0, and which in turn corresponds to a probability of approximately .02. Furthermore, $\exp(1.11) \approx 3.03$, indicating that for an increase of 1 in the judged strength of the inferential connection between a conditional's antecedent and its consequent, we may expect to see a close to 75 percent increase in the odds of that conditional being judged true. To explain further, this means that, for instance, going from a 1 in inference strength to a 2 will increase the probability of choosing "True" from 7 percent to 18 percent, and going from a 4 in inference strength to a 5 will increase that probability from 67 percent to 86 percent. In other words, truth judgments are strongly associated with perceived strength of inferential connectedness, in line with previous experimental results.

We also note that the model performs very well. It classifies 87 percent of the participants'

first and the group that had received the truth rating task first. There was never an indication of an order effect, and so we report analyses based on the pooled response sets.

[12] In giving 89 percent Highest Density Intervals, we are following McElreath (2018), who proposes 89 percent (rather than some "round" percentage, like 90 or 95) because it more directly reminds us of the arbitrary nature of such conventional thresholds. (Roughly, the 89 % HDI comprises those values of the relevant parameter which possess some minimal level of posterior credibility, such that their total probability adds up to 89 percent.)

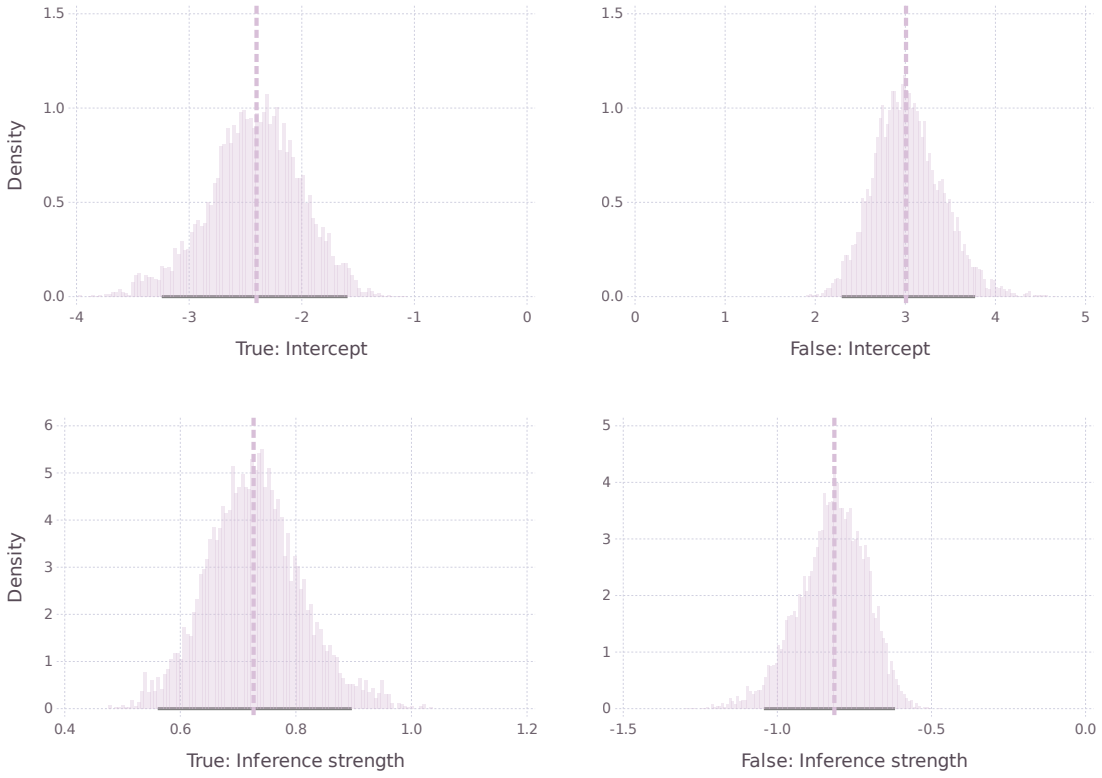

Figure 5: Posterior distributions for the parameters of the mixed-effects Bayesian multinomial regression model with truth responses (true/false/neither) as response variable and inference strength as fixed effect.

true/false judgments correctly, and its AUC value equals .91, which indicates exceptionally good discrimination (Hosmer, Lemeshow, & Sturdivant, 2013, p. 177).

For the true/false/neither responses, we conducted a mixed-effects multinomial logistic regression, choosing "Neither/nor" as the reference category. It had the same predictors as the previous model as well as the same priors. MCMC diagnostics did not raise any red flags for this model.

The median of the posterior distribution for the intercept for the "False" category was 3.03 (89 % HDI [2.29, 3.77]), that for the "Truth" category was −2.45 (89 % [−3.23, −1.57]). The medians of the posterior distributions for inference strength were −0.82 (89 % HDI [−1.04, −0.62]) for the "False" category and 0.73 (89 % HDI [0.58, 0.90]) for the "True" category. Figure 5 shows the full posterior distributions. The results indicate that if inference strength goes up by 1 point on a 7-point Likert scale, then we should expect the multinomial log-odds for a "False" response relative to a "Neither/nor" response to go down by 0.82 and the multinomial log-odds for a "True" response relative to a "Neither/nor" response to go up by 0.73.

Table 1 shows the confusion matrix for the multinomial logistic regression model. One easily verifies that the model has an overall accuracy of 75 percent, with most of the mistakes occurring for the "Neither/nor" responses, as one might also have expected given our earlier remarks on this category.

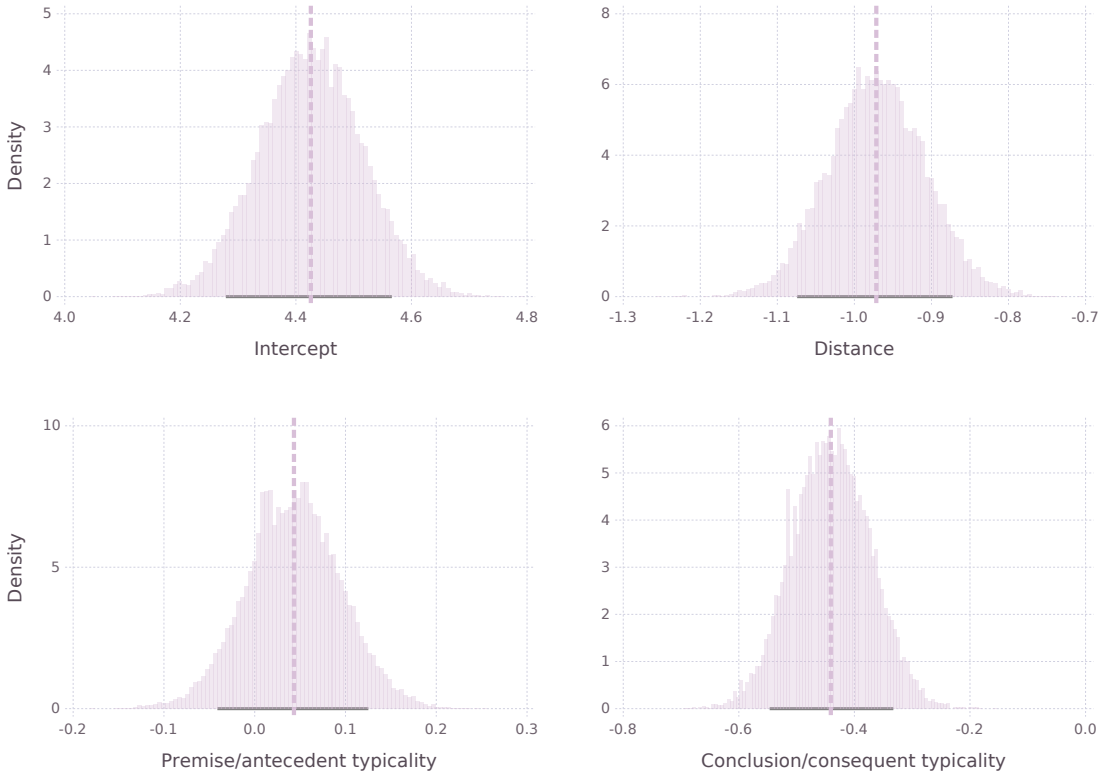

Figure 6: Posterior distributions for the parameters of the mixed-effects Bayesian linear model with perceived inference strength as response variable and distance, premise/antecedent typicality, and conclusion/consequent typicality as fixed effects.

4.2.2 Predicting perceived inference strength on the basis of distance and typicality

More central to the present study is the question whether perceived strength of the inferential connection between premise and conclusion (doubling as antecedent and consequent in the truth rating task) can be predicted on the basis of (i) the distance in our three-dimensional city-block space between the shapes designated in premise and conclusion (predictions 1 and 4, respectively), and (ii) the typicality (qua vase) of those same shapes (predictions 2 and 5, respectively).

We took *un*typicality to be measured by the city-block distance to the closest of the shapes that had been determined to be typical for vases in the second experiment of Douven (2016a). Thus, a distance of 0 to the closest such shape indicated maximum *typicality*. To facilitate interpretation, and to improve sampling efficiency, we centered all predictor variables on their

Table 1: Confusion matrix for the multinomial logistic regression model based on inference strength.

		predicted		
		true	false	neither
observed	true	258	45	8
	false	37	546	13
	neither	55	116	38

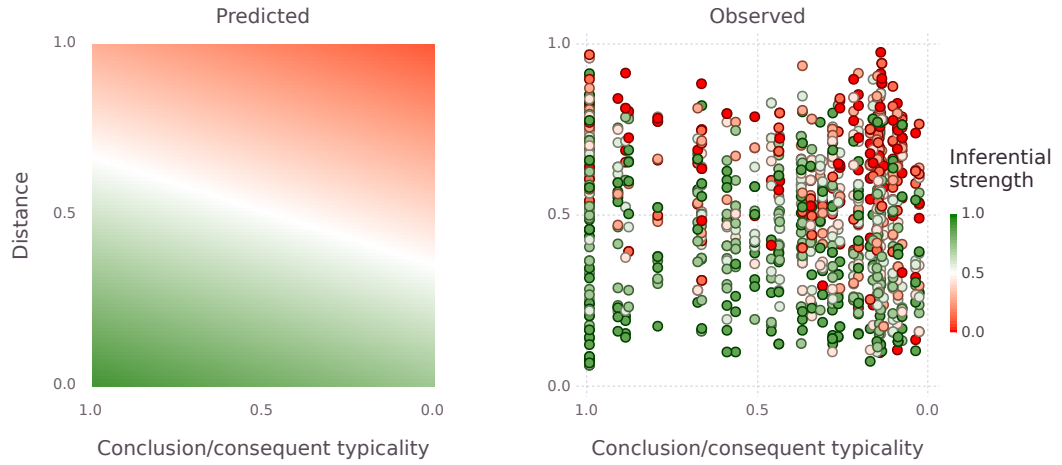

Figure 7: Plot of the mixed-effects Bayesian linear model (left) alongside the data (right).

mean and divided them by their standard deviation.

Using the `Turing.jl` package again, we conducted a mixed-effects Bayesian linear regression, with inference strength (as measured on a 7-point Likert scale) as response variable and with (scaled) distance between premise- and conclusion-shape in city-block space, typicality of premise-shape, and typicality of conclusion-shape as fixed effects. For all fixed effects, we also included per-participant intercepts and slopes as random effects. Priors were standard normal distributions for all fixed effects, a $\mathcal{N}(4, 1)$ distribution for the fixed effect intercept (i.e., centered on the midpoint of the response scale), $\mathcal{N}(0, 4)$ distributions truncated between 0 and infinity for the standard deviations in the fixed effects and fixed intercept, Cauchy$_+$(0, 2.5) distributions for the standard deviations in the random effects, and an LKJ(2) distribution for the correlation matrix of the random effects. Here, too, MCMC diagnostics indicated a sufficient mixing of the chains, sufficiently high bulk and tail effective sample size values, and an $\widehat{R}$ convergence diagnostic of 1.00 for every parameter.

The median of the posterior distribution for the intercept was 4.43 (89 % HDI [4.28, 4.57]) and those of the posterior distributions for distance, premise/antecedent typicality, and conclusion/consequent typicality were, respectively, −0.97 (89 % HDI [−1.07, −0.87]), 0.04 (89 % HDI [−0.04, 0.13]), and −0.44 (89 % HDI [−0.55, −0.33]). For the full posterior distributions, see Figure 6.

Thus, city-block distance between premise/antecedent shape and conclusion/consequent shape is strongly negatively associated with perceived inference strength: all else being equal, an increase in that distance by one standard deviation is associated with a lowering of perceived inference strength by approximately one point on a 7-point Likert scale. By contrast, there is a moderately strong positive association between perceived inference strength and conclusion/consequent typicality: all else being equal, an increase in conclusion/consequent typicality (which, note, is a decrease in distance from nearest prototype) by one standard deviation is associated with an increase in perceived inference strength by almost half a point on a 7-point Likert scale. Finally, there is no meaningful association between premise/antecedent typicality and perceived inference strength.

Figure 7 plots the model alongside the data, where the axes in both plots correspond to the two variables meaningfully connected with inferential strength, and with color indicating

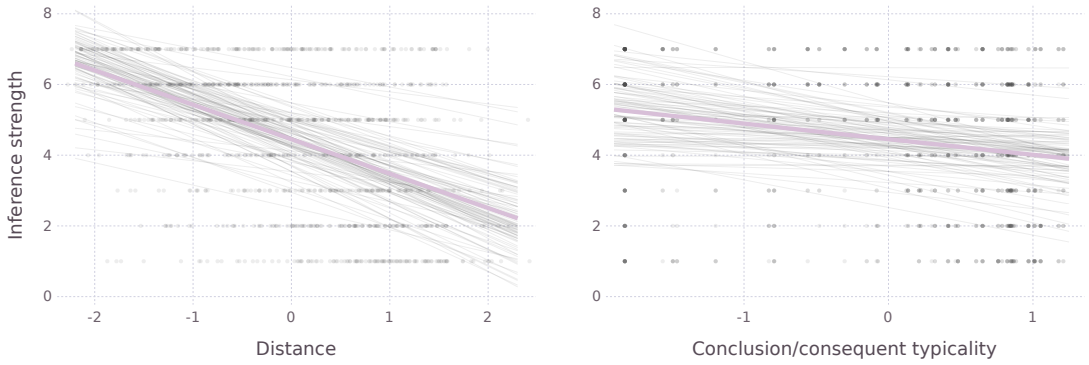

Figure 8: Marginal effects plots of the main predictors from the mixed-effects Bayesian linear model. The thin gray lines show the predictions of the individual random effects from all participants.

inferential strength. For ease of interpretation, both distance and conclusion/consequent typicality were normalized, the zero point corresponding with the hypothetical case of the premise/antecedent and conclusion/consequent shape being identical, respectively, with the conclusion/consequent shape being a prototypical vase shape, and the one point with the maximum distance in city-block space assumed by a pair of shapes, respectively, the maximum distance from any of the prototypical vase shapes. Figure 8 plots the marginal effects of distance and conclusion/consequent typicality.

The strong negative association between inference strength and distance is clear support for the main hypothesis that the geometry of our concepts explains which proximity-influenced inferences we are, and which we are not, willing to make (prediction 1). We saw that, at least for the closely related category-based inductive inferences, Osta-Vélez and Gärdenfors (2020) also postulate a role for typicality in explaining inference, which is why we incorporated typicality claims in our main hypothesis (prediction 2). Our results might seem to offer *partial* support for that hypothesis, given that at least conclusion/consequent typicality had a clear positive impact on perceived inference strength, even if premise/antecedent typicality did not. But this finding is to be interpreted with some caution, especially in view of the literature on belief bias, according to which people are more inclined to infer a conclusion they already find plausible, independent of the argument given for that conclusion. For it is reasonable to assume that the more typical a shape is for, say, vases, the more people will be inclined to believe it is a vase, independently of how the shape designated in the premise is classified according to that premise. But then, because of belief bias, people will be more likely to say that the conclusion follows, regardless of which shape figures in the antecedent. We note that belief bias has been recorded for inductive inference (reviewed in Dube, Rotello, & Heit, 2010) as well as informal inference (Thompson & Evans, 2012).[13] We will take this up again in the next section.

4.2.3 Predicting truth ratings on the basis of distance and typicality

Finally, we turn to our second hypothesis, and so the question of whether responses in the truth rating task could be predicted purely on the basis of information about the city-block space in which the shapes referred to in the materials can be represented, specifically, the distance in

[13] Some might be tempted to speculate that typicality actually *explains* belief bias. We do not believe there would be much merit to that explanation, however, given that the belief bias effect is found across a broad range of argument types, including ones for which the notion of conclusion/consequent typicality makes little sense.

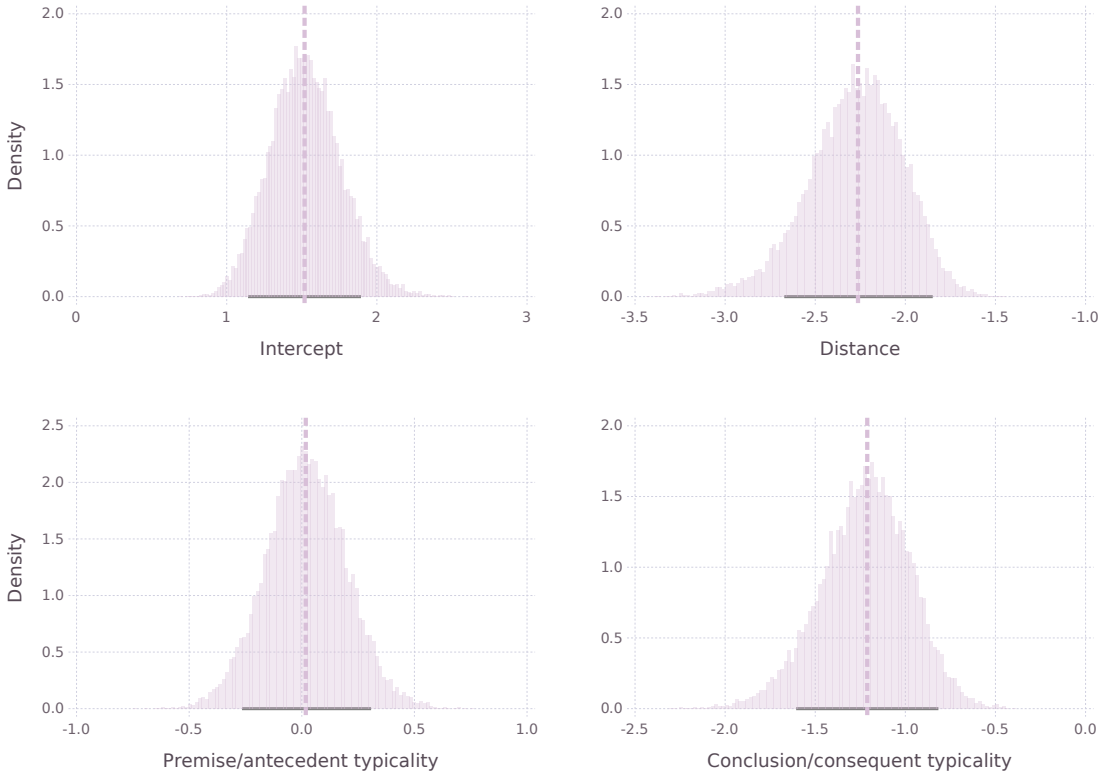

Figure 9: Posterior distributions for the parameters of the mixed-effects Bayesian logistic model with truth responses (true/false) as response variable and distance, premise/antecedent typicality, and conclusion/consequent typicality as fixed effects.

that space between the objects designated in a conditional, and the typicality of those objects, *qua* vase (predictions 4 and 5). This is also our opportunity to disentangle predictions 5a and 5b. Here, too, we first look at true/false responses, excluding the neither responses, and then at true/false/neither responses.

For the true/false responses, we again conducted a Bayesian mixed-effects logistic regression, which had the said responses as a binary response variable and the same predictors as before were used in the linear regression described above: distance between shapes, premise/antecedent typicality, and conclusion/consequent typicality. Also as before, we added per-participant random intercepts and slopes for all predictors. We used the same priors as in the linear model. MCMC diagnostics again gave no reason for concern.

The posterior distribution for the intercept had a median of 1.52 (89 % HDI [1.14, 1.89]) and those for distance, premise/antecedent typicality, and conclusion/consequent typicality had medians of −2.26 (89 % HDI [−2.67, −1.85]), 0.02 (89 % HDI [−0.27, 0.31]), and −1.21 (89 % HDI [−1.61, −0.82]), respectively. Figure 9 shows the full posterior distributions. Figure 10 plots probability of a "True" response predicted by the model on the basis of the two main predictors; for comparison, it also shows the observed responses for combinations of distance and conclusion/consequent typicality as they occurred in our materials.

These findings mean the following: because $\exp(-2.26) \approx 0.1$, the odds for a conditional to be judged true if the distance between the shapes it refers to increases by one standard deviation are lowered by a factor of 10, all else being equal. On the other hand, if the typicality

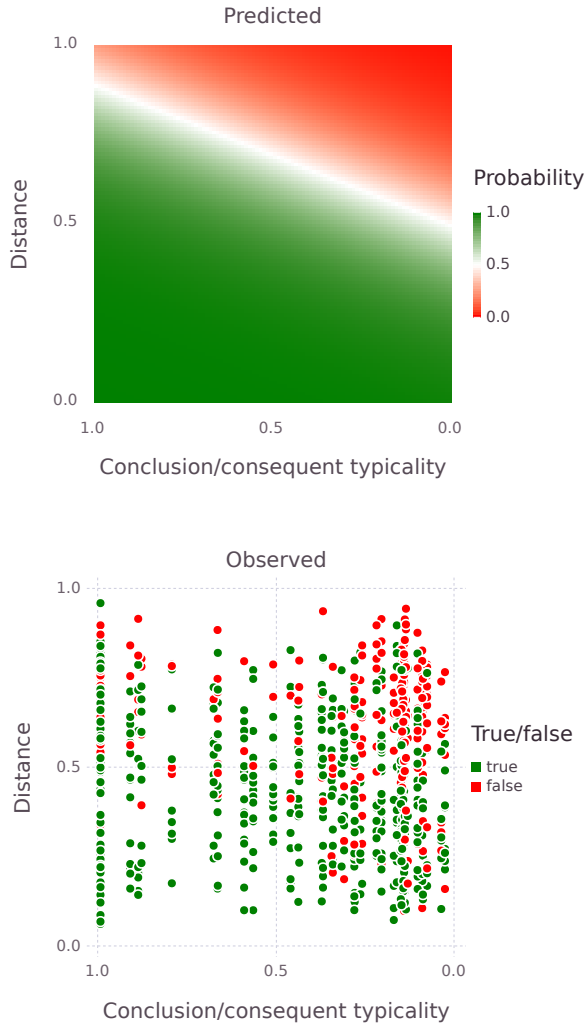

Figure 10: Plot of the mixed-effects Bayesian logistic model (top) together with the data (bottom).

of the shape a conditional's consequent refers to increases by one standard deviation, the odds of that conditional being judged true go up by about 77 percent (given that $\exp(1.21) \approx 3.35$), all else being equal. Thus, we supported prediction 5b in preference to prediction 5a.

While this model did not perform quite as well as the earlier model that predicted true/false ratings on the basis of inference strength ratings, it still has a more than satisfactory performance, classifying 77 percent of the participants' true/false judgments correctly, and having an AUC value of .81, which according to Hosmer, Lemeshow, and Sturdivant (2013, p. 177) still indicates excellent discrimination.

For the true/false/neither responses, we conducted a Bayesian multinomial logistic regression, with truth ratings (now including "Neither/nor" responses) as the response variable. Predictors and priors were as in the binary logistic regression described in Section 4.2.1. We took "Neither/nor" as the reference category. The MCMC diagnostics gave no cause for concern.

The median of the posterior distribution for the intercept for the "False" category was −0.28 (89 % HDI [−0.63, 0.05]), that of the posterior distribution for the intercept for the "Truth" category was 1.22 (89 % [0.90, 1.54]). The medians of the posterior distributions for

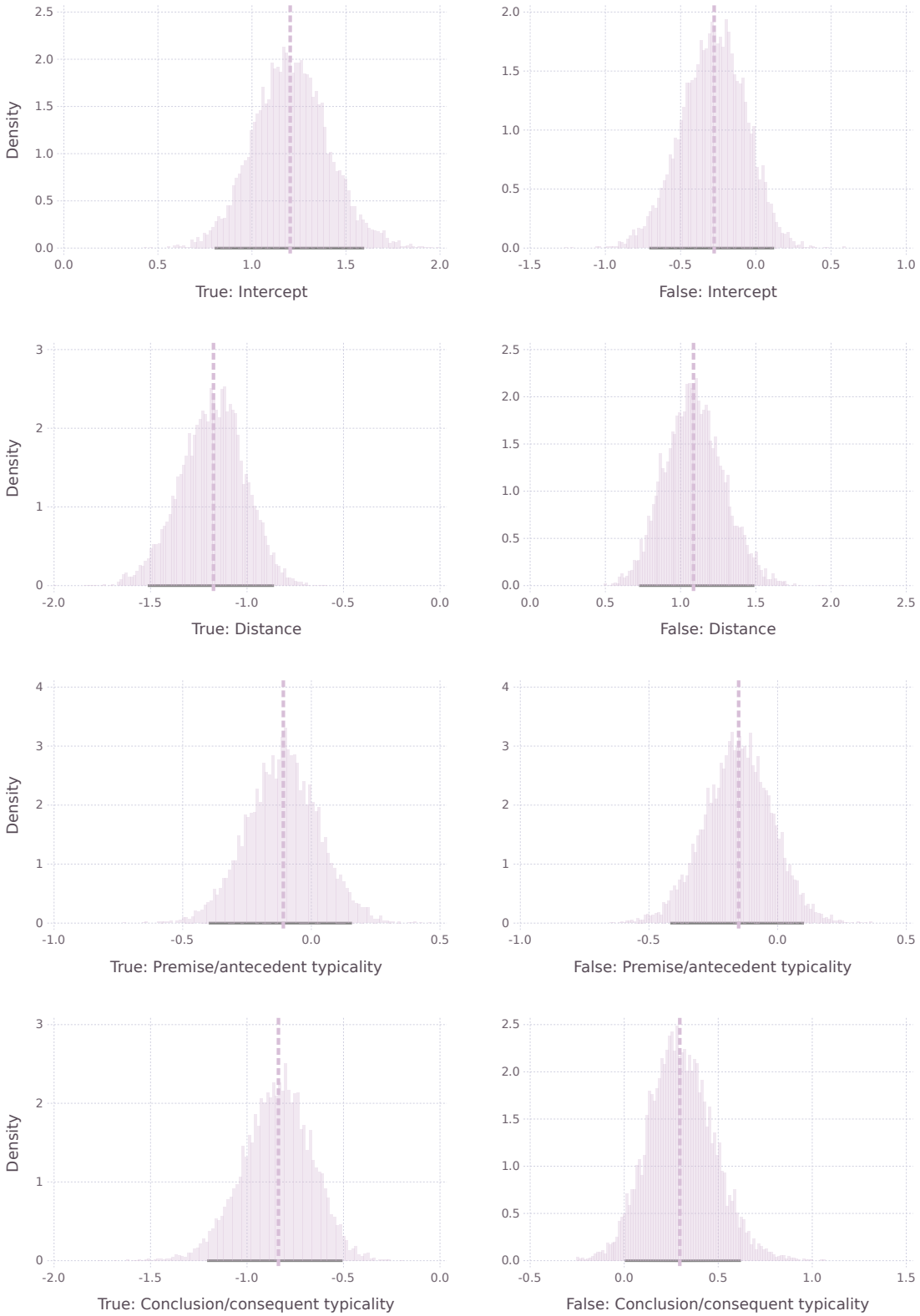

Figure 11: Posterior distributions for the parameters of the mixed-effects Bayesian multinomial logistic model with truth responses (true/false/neither) as response variable and distance, premise/antecedent typicality, and conclusion/consequent typicality as fixed effects.

Table 2: Confusion matrix for the multinomial logistic regression model based on distance, premise/antecedent typicality, and conclusion/consequent typicality.

		predicted		
		true	false	neither
observed	true	257	51	3
	false	32	553	11
	neither	50	76	83

distance, premise/antecedent typicality, and conclusion/consequent typicality were, for the "False" category, 1.09 (89 % HDI [0.78, 1.41]), −0.15 (89 % HDI [−0.37, 0.05]), and 0.30 (89 % HDI [0.02, 0.58]), respectively, and for the "True" category, −1.18 (89 % HDI [−1.43, −0.90]), −0.11 (89 % HDI [−0.33, 0.12]), and −0.84 (89 % HDI [−1.13, −0.56]), respectively. See Figure 11 for the full posterior distributions.

Most importantly, this means that (i) if the distance between shapes increases by one standard deviation, the multinomial log-odds for a "False" response relative to a "Neither/nor" response is expected to increase by 1.09, holding the other variables in the model constant; (ii) if the distance between shapes increases by one standard deviation, the multinomial log-odds for a "True" response relative to a "Neither/nor" response is expected to decrease by 1.18, also holding all else constant; (iii) if the typicality of the conclusion/consequent shape increases by one standard deviation (i.e., the distance between the shape and the nearest vase prototype decreases by one standard deviation), the multinomial log-odds for a "False" response relative to a "Neither/nor" response is expected to decrease by 0.3, all else being equal; and finally (iv) if the typicality of the conclusion/consequent shape increases by one standard deviation, the multinomial log-odds for a "True" response relative to a "Neither/nor" response is expected to increase by 0.84, all else being equal. Observe that, here too, we find no noteworthy association between truth ratings and premise/antecedent typicality.

As for performance, Table 2 gives the confusion matrix for the multinomial logistic regression model. We see that the model predictions match the observations fairly well, with an overall accuracy of 80 percent. Most of the mistakes again occur for the participants' "Neither/nor" responses, which is not surprising given the mixed nature of this category, as explained previously. Figure 12 compares the observed responses with the predictions made by the model.

5 General discussion

Our focus in this paper was on an under-studied subtype of non-deductive arguments, what borrowing an expression from Carnap (1980) we called "proximity-influenced arguments," that is, arguments which project a property from one object onto another, on account of the objects being similar to each other. In line with recent work on the related type of category-based inductive arguments, we hypothesized that the strength of proximity-influenced arguments would depend on how similar are the objects designated in the premise and the conclusion, and on how typical they are for the projected property. From this hypothesis together with Inferentialism, it follows that those same factors also predict endorsement rates of conditionals whose component parts are connected by a proximity-influenced inferential link.

Both hypotheses were further specified within the conceptual spaces framework, to the effect that the aforementioned factors would be accurate predictors if interpreted in terms of

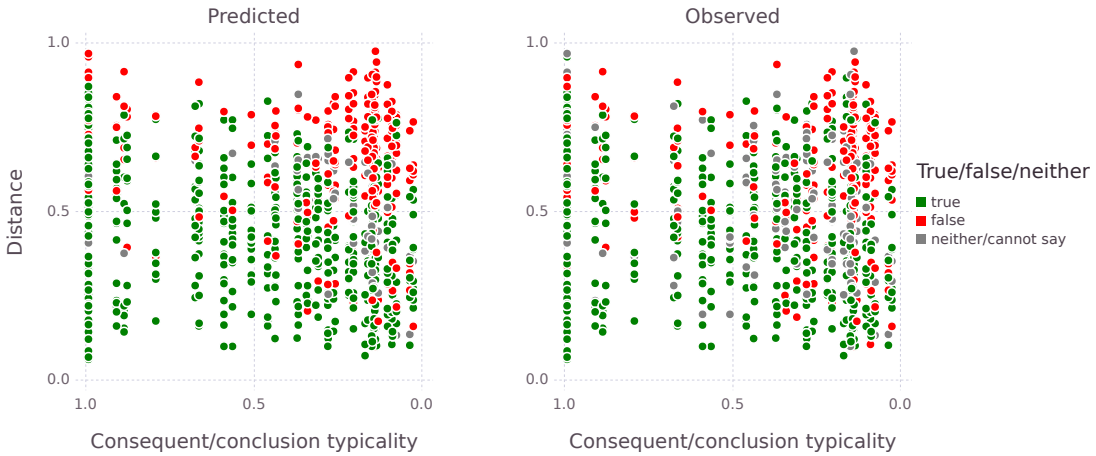

Figure 12: Plot of the predictions from the multinomial logistic regression model (left) alongside the data (right).

distances in the relevant conceptual space, which for our materials was the shape space from Douven (2016a). In particular, we hypothesized that a proximity-influenced argument would be perceived to be stronger (i) the closer to each other the objects involved in the argument are, as represented in that space, ceteris paribus; (ii) the closer the object designated in the premise is to the prototypical region of the projected property, ceteris paribus; and (iii) the closer the object designated in the conclusion is to the prototypical region of the projected property, ceteris paribus. The hypothesis concerning truth ratings of conditionals with proximity-influenced inferential links was cashed out in parallel fashion.

Our study provided strong support for corresponding parts of both hypotheses: premise/antecedent–conclusion/consequent similarity was strongly associated with perceived argument strength, respectively, probability of truth, as was—to a slightly lesser extent—conclusion/consequent typicality; there was no meaningful role for premise/antecedent typicality. As noted, however, in interpreting the finding regarding conclusion/consequent typicality, we have to reckon with the possibility that this finding at least partly betokens a belief bias effect. This is well in line with HIT, the psychological counterpart of Inferentialism, and with previously recorded findings (Douven et al., 2018).

To our knowledge, we have provided the first evidence that the conceptual spaces framework can be fruitfully mustered for explaining certain patterns of non-deductive reasoning as well as for the evaluation of certain types of conditionals. Our study could serve as a template for testing other types of non-deductive reasoning in this framework. The study follows the general methodology for investigating inferences based on concept structures proposed in Osta-Vélez and Gärdenfors (2020, Sect. 6).

Perhaps the most obvious follow-up research would turn to category-based induction directly and try and test Osta-Vélez and Gärdenfors' (2020) proposal that was discussed in Section 3. These authors illustrate their proposal by means of a bird space and a mammal space, noting however that neither bird space nor mammal space is actually available for researchers to work with. More generally, at the time we lack the precise knowledge of the geometry and topology of a conceptual space that has a sufficiently fine-grained structure to allow a proper empirical validation of Osta-Vélez and Gärdenfors' proposal.

Apart from case-based induction, the methodology could be applied to other forms of non-deductive reasoning. A first example is that Osta-Vélez and Gärdenfors (submitted) have proposed that non-monotonic reasoning—reasoning based on more or less strong default assumptions or background knowledge—can be analyzed in terms of distances from prototypes in conceptual spaces. In contrast to other accounts of non-monotonic reasoning, the model they propose generates predictions concerning the strength of the arguments. A methodology similar to that of the present article could be used to test this model empirically.

Another area would be the validity of generics, such as "French people like wine" and "Tigers are striped." This is an area that has been much discussed in philosophy, linguistics, and psychology. Gärdenfors and Osta-Vélez (submitted) have proposed a model of generics that presents them as expectations of various strengths that can be added to general background knowledge in reasoning. This model is also based on distances to prototypes in conceptual spaces. Differences of strength have not been studied in the literature, but it seems obvious that a generic such as "Elephants have trunks" is judged to be more valid than "Elephants are grey" since an exception to the latter would be more similar to the prototypical elephant than an exception to the former.

These potential applications of the present methodology to other areas of similarity-based non-deductive reasoning would require the identification of conceptual spaces with a richer structure than our shape space. Such spaces might not be easily attainable. Consider, for instance, what might at first appear a rather straightforward exercise, to wit, fine-graining color space. We already have the CIELab and CIELuv spaces and also know, in those spaces, the prototypical regions for Berlin and Kay's (1969) eleven basic color concepts (Douven, 2019a). For a seemingly modest start, we might try to fine-grain BLUE, say, carving it up into regions corresponding to, perhaps, AQUAMARINE, AZURE, TEAL, TURQUOISE, and so on. We could further try to coarse-grain color space into warm and cold colors, also determining which color shades people consider typically warm and which typically cold. The resulting space would certainly be rich enough to test Osta-Vélez and Gärdenfors' proposal in detail. Note, however, how hard it would be in practice to acquire the requisite data. While all of the present authors are familiar with the fact that aquamarine and azure are shades of blue, not all of us are able to reliably identify these shades. More generally, in a large-scale study on color perception, Jraissati and Douven (2018) found a large variation in participants' use of non-basic color terms, even though responses were quite consistent for basic color terms. One can imagine how difficult it would be, as a result, to obtain reliable information about the locations in color space of the prototypes, or prototypical regions, of aquamarine, turquoise, and so on.

Even if challenging, the effort to make conceptual spaces more broadly available for investigating inference patterns is well worth making. It is a common observation among philosophers that, while we have strict norms for deductive reasoning, such norms are still missing for non-deductive forms of reasoning (e.g., Maher, 2001; Bartha, 2010; Douven, 2021c). But at least for those non-deductive inferences that can be represented in conceptual spaces, in the way seen in this paper, norms of correctness may be forthcoming. Although, in principle, any Voronoi tessellation on a similarity space yields a conceptual space, Gärdenfors (2000) emphasized early on that we are only interested in those tessellations that produce *natural* concepts, that is, concepts of the kind that may figure in our thinking and communication. His early proposal for how to delineate natural from non-natural concepts invoked a kind of rationality principle (or design principle), but Douven and Gärdenfors (2020) pointed out that, on its own, that principle is too weak to yield the desired result (see also Douven, 2019b). To fix this problem, Douven and Gärdenfors state a number of additional rationality principles that—they

argue—together do suffice to yield carvings-up of conceptual spaces that result in truly natural concepts. In a nutshell, their claim is that natural concepts are concepts represented by the cells of an optimally (i.e., most rationally) partitioned similarity space.

With norms for the correct design of conceptual spaces in place, we obtain a normative account of similarity-based inferences almost for free. For instance, a proximity-influenced argument could be said to be valid to the extent that the objects designated in its premise and conclusion lie close to each other in the conceptual space in which the projected property is represented, provided that conceptual space is optimally partitioned. The definition of validity for category-based inductive arguments would be essentially the same. One could consider giving typicality a place in those definitions, but while we see no good reason for doing that, we want to leave this open for debate for now.

A further avenue for future research is suggested by a remark we made in Section 2.1 about competing approaches to conceptualization. While, as we said, we are not aware of accounts of similarity-based reasoning built upon those approaches, it would be worth trying to develop such accounts and compare them with the proposal examined in this paper. Here, we also call upon proponents of the other approaches to conceptualization.

To end, we would like to mention a potential limitation of the present work. The study we reported was conducted online. In psychology and the social sciences, online studies have gained widespread popularity, partly because crowdsourcing services such as Amazon's Mechanical Turk and Prolific make it easy to collect vast amounts of responses in a matter of days or even hours. Collecting the same number of responses in a laboratory would often be unaffordable or at least highly impractical. Online studies done via crowdsourcing platforms have the inherent advantage of accessing a general audience rather than the first-year psychology students which are the typical population of laboratory studies. Moreover, the specific platform we used, Prolific, has a raft of measures in place to enhance response validity, and has been found to be superior in data quality to other platforms such as MTurk and CrowdFlower (Peer et al., 2017). These advantages are almost impossible to replicate in the laboratory. But even though in general online studies have become the norm in many domains of cognitive psychology, one might still have concerns over their use for perception studies and similar types of studies involving visual stimuli, that previously were conducted under strictly controlled viewing conditions. While one can incorporate all sorts of attention checks in an online survey—as we did in the one used for the present work—it is not possible to achieve the kind of control on viewing conditions that one typically has in a laboratory. It is to be stressed, however, that a number of replication studies especially concerned with color perception—an area in which viewing conditions are probably still more important than in the research we reported, which only involved shape perception—obtained virtually the same results that had earlier been obtained in a specialized laboratory (Moroney, 2003; Mylonas & MacDonald, 2010; Mylonas, Paramei, & MacDonald, 2014). While the outcomes of these studies are encouraging, we do hope that we or other researchers will be able to further confirm our hypotheses by rerunning our online experiment, and running related ones (e.g., in the context of the aforementioned follow-up projects), in a controlled environment.[14]

References

Abdi, H. & Williams, L. J. (2010). Principal component analysis. *WIREs Computational Statistics*,

[14]We are grateful to two anonymous referees for valuable feedback and to Mike Oaksford for helpful editorial guidance. Thanks also to an audience at Ruhr University Bochum for stimulating questions and discussion.

2, 433–459.

Albright, A. & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, *90*, 119–161.

Alchourrón, C. E., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, *50*, 510–530.

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*, 527–535.

Bartha, P. (2010). *By Parallel Reasoning*. Oxford: Oxford University Press.

Bellmund, J. L. S., Gärdenfors, P., Moser, E. I., & Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science*, *362*, 6415.

Berlin, B. & Kay, P. (1969). *Basic Color Terms*. Stanford CA: CSLI Publications.

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*, 65–98.

Borg, I. & Groenen, P. (2010). *Modern Multidimensional Scaling* (2nd ed.). New York: Springer.

Carnap, R. (1952). Meaning postulates. *Philosophical Studies*, *3*, 65–73.

Carnap, R. (1980). A basic system of inductive logic II. In R. C. Jeffrey (ed.), *Studies in Inductive Logic and Probability* (pp. 7–155). Berkeley CA: University of California Press.

Corner, A., Hahn, U., & Oaksford, M. (2011). The psychological mechanism of the slippery slope argument. *Journal of Memory and Language*, *64*, 133–152.

Dawdy-Hesterberg, L. G. & Pierrehumbert, J. B. (2014). Learnability and generalisation of Arabic broken plural nouns. *Language, Cognition and Neuroscience*, *29*, 1268–1282.

Decock, L. & Douven, I. (2011). Similarity after Goodman. *Review of Philosophy and Psychology*, *2*, 61–75.

Decock, L. & Douven, I. (2014). What is graded membership? *Noûs*, *48*, 653–682.

Douven, I. (2016a). Vagueness, graded membership, and conceptual spaces. *Cognition*, *151*, 80–95.

Douven, I. (2016b). *The Epistemology of Indicative Conditionals*. Cambridge: Cambridge University Press.

Douven, I. (2017). How to account for the oddness of missing-link conditionals. *Synthese*, *194*, 1541–1554.

Douven, I. (2019a). Putting prototypes in place. *Cognition*, *193*, 104007, doi: 10.1016/j.cognition.2019.104007.

Douven, I. (2019b). The rationality of vagueness. In R. Dietz (ed.), *Vagueness and Rationality* (pp. 115–134). New York: Springer.

Douven, I. (2021a). Implicatures and naturalness. In S. Löbner, T. Gamerschlag, T. Kalenscher, M. Schrenk, & H. Zeevat (eds.), *Concepts, Frames and Cascades in Semantics, Cognition and Ontology* (pp. 143–163). Cham: Springer.

Douven, I. (2021b). Fuzzy concept combination: An empirical study. *Fuzzy Sets and Systems*, *407*, pp. 27–49.

Douven, I. (2021c). *The Art of Abduction*. Cambridge MA: MIT Press, in press.

Douven, I. & Decock, L. (2017). What verities may be. *Mind*, *126*, 386–428.

Douven, I., Decock, L., Dietz, R., & Égré, P. (2013). Vagueness: A conceptual spaces approach. *Journal of Philosophical Logic*, *42*, 137–160.

Douven, I., Elqayam, S., Singmann, H., & van Wijnbergen-Huitink, J. (2018). Conditionals and inferential connections: A hypothetical inferential theory. *Cognitive Psychology*, *101*, 50–81.

Douven, I., Elqayam, S., Singmann, H., & van Wijnbergen-Huitink, J. (2020). Conditionals

and inferential connections: Toward a new semantics. *Thinking & Reasoning*, *26*, 311–351.
Douven, I. & Gärdenfors, P. (2020). What are natural concepts? A design perspective. *Mind & Language*, *35*, 313–334.
Douven, I., Wenmackers, S., Jraissati, Y., & Decock, L. (2017). Measuring graded membership: The case of color. *Cognitive Science*, *41*, 686–722.
Dube, C., Rotello, C. M., & Heit, E. B. (2012). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review*, *110*, 831–863.
Evans, J. St. B. T. (2006). The heuristic–analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review*, *13*, 378–395.
Evans, J. St. B. T. (2007). *Hypothetical Thinking: Dual Processes in Reasoning and Judgement*. Hove UK: Psychology Press.
Evans, J. St. B. T., Barston, J. L., & Pollard, P. B. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, *11*, 295–306.
Fairchild, M. D. (2013). *Color Appearance Models*. Hoboken NJ: Wiley.
Gärdenfors, P. (2000). *Conceptual Spaces*. Cambridge MA: MIT Press.
Gärdenfors, P. (2014). *The Geometry of Meaning*. Cambridge MA: MIT Press.
Gärdenfors, P. & Osta-Vélez, M. (submitted). Generics as expectations.
Goodman, N. (1972). Seven strictures on similarity. In his *Problems and Projects* (pp. 437–446). Indianapolis/New York: Bobbs–Merrill.
Hahn, U. & Nakisa, R. C. (2000). German inflection: Single route or dual route? *Cognitive Psychology*, *41*, 313–360.
Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition*, *65*, 137–165.
Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science*, *31*, 355–384.
Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Hoboken NJ: Wiley.
Hout, M. C., Papesh, M. H., & Goldinger, S. D. (2013). Multidimensional scaling. *WIREs Cognitive Science*, *4*, 93–103.
Jraissati, Y. & Douven, I. (2018). Delving deeper into color space. *i-Perception*, *9*, 1–27, doi: 10.1177/2041669518792062.
Kamp, H. & Partee, B. (1995). Prototype theory and compositionality. *Cognition*, *57*, 129–191.
Kneale, W. & Kneale, M. (1962). *The Development of Logic*. Oxford: Oxford University Press.
Kruschke, J. K. (2015). *Doing Bayesian Data Analysis* (2nd ed.). Amsterdam: Elsevier.
Krzyżanowska, K. H. (2015). *Between "If" and "Then."* Doctoral dissertation. University of Groningen.
Krzyżanowska, K. H., Collins, P. J., & Hahn, U. (2017). Between a conditional's antecedent and its consequent: Discourse coherence vs. probabilistic relevance. *Cognition*, *164*, 199–205.
Krzyżanowska, K. H., Collins, P. J., & Hahn, U. (2021). True clauses and false connections. *Journal of Memory and Language*, in press.
Krzyżanowska, K. H. & Douven, I. (2018). Missing-link conditionals: pragmatically infelicitous or semantically defective? *Intercultural Pragmatics*, *15*, 191–211.
Krzyżanowska, K. H., Wenmackers, S., & Douven, I. (2014). Rethinking Gibbard's river-boat argument. *Studia Logica*, *102*, 771–792.
Lycan, W. G. (2001). *Real Conditionals*. Oxford: Oxford University Press.
Maher, P. (2001). Probabilities for multiple properties: The models of Hesse and Carnap and

Kemeny. *Erkenntnis*, *55*, 183–216.
Mackie, J. L. (1973). *Truth, Probability and Paradox: Studies in Philosophical Logic*. Oxford: Oxford University Press.
McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton FA: CRC Press.
Mellor, D. H. & Bradley, R. (2021). Conditionals: Truth, safety, and success. *Mind & Language*, in press.
Mill, J. S. (1843/1872). *A System of Logic: Ratiocinative and Inductive* (8th ed.). London: Longmans, Green, Reader, & Dyer.
Mirabile, P. & Douven, I. (2020). Abductive conditionals as a test case for inferentialism. *Cognition*, *200*, 104232, doi: 10.1016/j.cognition.2020.104232.
Moroney, N. (2003). Unconstrained web-based color naming experiment. In R. Eschbach & G. G. Marcu (eds.), *Color Imaging VIII: Processing, Hardcopy, and Applications* (Proc. SPIE, vol. 5008, pp. 36–46). SPIE.
Mylonas, D. & MacDonald, L. (2010). Online colour naming experiment using Munsell samples. In *Proceedings of the 5th European Conference on Colour in Graphics, Imaging, and Vision* (pp. 27–32). Society for Imaging Science and Technology.
Mylonas, D., Paramei, G. V., & MacDonald, L. (2014). Gender differences in colour naming. In W. Anderson, C. P. Biggam, C. Hough, & C. Kay (eds.), *Colour Studies: A Broad Spectrum* (pp. 225–239). Philadelphia PA: John Benjamins Publishing Company.
Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.
Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 87–108.
Nosofsky, R. M. (1989). Further tests of an exemplar–similarity approach to relating identification and categorization. *Perception and Psychophysics*, *45*, 279–290.
Nosofsky, R. M. & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 924–940.
Oaksford, M. & Chater, N. (2010). Causation and conditionals in the cognitive science of human reasoning. *The Open Psychology Journal*, *3*, 105–118.
Oaksford, M. & Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. *Thinking & Reasoning*, *19*, 346–379.
Oaksford, M. & Chater, N. (2014). Probabilistic single function dual process theory and logic programming as approaches to non-monotonicity in human vs. artificial reasoning. *Thinking & Reasoning*, *20*, 269–295.
Oaksford, M. & Chater, N. (2017). Causal models and conditional reasoning. In M. R. Waldmann (ed.), *The Oxford Handbook of Causal Reasoning* (pp. 327–346). Oxford: Oxford University Press.
Oaksford, M. & Chater, N. (2020). Integrating causal Bayes nets and inferentialism in conditional inference. In S. Elqayam, I. Douven, J. St. B. T. Evans, & N. Cruz (eds.), *Logic and Uncertainty in the Human Mind* (pp. 116–132). London: Routledge.
Okabe, A., Boots., B., Sugihara, K., & Chiu, S. N. (2000). *Spatial Tessellations* (2nd ed.). New York: Wiley.
Osherson, D., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *32*, 185–200.

Osta-Vélez, M. & Gärdenfors, P. (2020). Category-based induction in conceptual spaces. *Journal of Mathematical Psychology*, *96*, 102357, doi: 10.1016/j.jmp.2020.102357.

Osta-Vélez, M. & Gärdenfors, P. (submitted). Nonmonotonic reasoning, expectation orderings, and conceptual spaces.

Paris, J. B. & Vencovská, A. (2017). Combining analogical support in pure inductive logic. *Erkenntnis*, *82*, 401–419.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163.

Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base rates: Both neglected and intuitive. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 544–554.

Ramsey, F. P. (1929/1990). General propositions and causality. In his *Philosophical Papers*, edited by D. H. Mellor (pp. 145–163). Cambridge: Cambridge University Press.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *4*, 328–350.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (eds.), *Cognition and Categorization* (pp. 27–48). Hillsdale NJ: Erlbaum.

Rostworowski, W., Pietrulewicz, N., & Będkowski, M. (2021). Conditionals and specific links: An experimental study. *Synthese*, in press.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, *1*, 54–87.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Simon, H. A. (1982). *Models of Bounded Rationality*. Cambridge MA: MIT Press.

Skovgaard-Olsen, N., Kellen, D., Hahn, U., & Klauer, K. C. (2019). Norm conflicts and conditionals. *Psychological Review*, *126*, 611–633.

Skovgaard-Olsen, N., Kellen, D., Krahl, H., & Klauer, K. C. (2017). Relevance differently affects the truth, acceptability, and probability evaluations of "and", "but", "therefore", and "if–then". *Thinking & Reasoning*, *23*, 449–482.

Sloman, S. (1993). Feature-based induction. *Cognitive Psychology*, *25*, 231–280.

Stalnaker, R. (1975). Indicative conditionals. *Philosophia*, *5*, 269–286.

Stalnaker, R. (1979). Anti-essentialism. *Midwest Studies in Philosophy*, *4*, 343–355.

Stewart, A. J., Singmann, H., Haigh, M., Woods, J. S., & Douven, I. (2021). Tracking the eye of the beholder: Is explanation subjective? *Journal of Cognitive Psychology*, in press.

Thompson, V. A. & Evans, J. St. B. T. (2012). Belief bias in informal reasoning. *Thinking & Reasoning*, *18*, 278–310.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–354.

van Fraassen, B. C. (1967). Meaning relations among predicates. *Noûs*, *1*, 160–179.

van Rooij, R. & Schulz, K. (2019). Conditionals, causality and conditional probability. *Journal of Logic, Language and Information*, *28*, 55–71.

Verheyen, S. & Égré, P. (2018). Typicality and graded membership in dimensional adjectives. *Cognitive Science*, *42*, 2250–2286.

Vidal, M. & Baratgin, J. (2017). A psychological study of unconnected conditionals. *Journal of Cognitive Psychology*, *29*, 769–781.

Vigo, R. & Allen, C. (2009). How to reason without words: Inference as categorization. *Cognitive Processing*, *10*, 77–88.