The Sorites Paradox in Psychology

Paul Egré, David Ripley, Steven Verheyen

THE SORITES PARADOX IN PSYCHOLOGY

PAUL ÉGRÉ, DAVID RIPLEY, STEVEN VERHEYEN

Abstract. This chapter examines some aspects of the influence of the sorites paradox in psychology. Section 1 starts out with a brief discussion of the analysis of slippery slope arguments in the psychology of reasoning, to introduce the relevance of probabilistic considerations in that domain. We then devote most of this chapter to the analysis in psychophysics and in the psychology of concepts of the complex relationship between discrimination and categorization for items that differ very little. Section 2 emphasizes the centrality of probabilistic modeling to represent the way in which small differences between stimuli affect decisions of membership under a common category. Section 3 focuses on experimental data concerning unordered transitions between prototypes, then section 4 looks at data concerning ordered transitions between prototypes (dynamic sorites).

This chapter examines some areas of theoretical and experimental psychology in which the sorites paradox has had an influence or has been an object of study. Our goal is to show not only different manifestations of the sorites in psychology, but also how psychological modeling and behavioral data can cast light on the puzzle raised by the paradox.

The first aspect we consider concerns the *psychology of reasoning and argumentation*. Sorites arguments are often conflated with 'slippery slope arguments', typically used *a contrario* to argue that a line should be drawn at a specific location of a vague domain, on pain of reaching an undesirable or absurd outcome, or alternatively that no line can be drawn at all. In Section 1 we start out with a brief history and overview of work done on slippery slope arguments, to highlight that such arguments are not intrinsically wrong: fundamentally they are inductive arguments, whose acceptability depends on the strength of the relation between the antecedent and the consequent of their conditional premises, and on the utility attached to specific outcomes. As such, slippery slope arguments tend to be handled in a probabilistic framework.

The second and more significant area of influence we consider, concerns the study of similarity in *psychophysics* and in the *psychology of concepts*. The main premise of a sorites argument involves the notion of sufficient similarity between objects, and states that if two objects are sufficiently similar, they must produce identical judgments as to whether some property applies or not. Section 2 presents some influential accounts of the relation between discrimination and categorization in psychophysics, and underscores the centrality of probabilistic modeling to deal with sorites-susceptible predicates quite generally.

We distinguish, following Raffman (1994) and Dzhafarov and Dzhafarov (2012), two versions of the main premise of the sorites, one pertaining to discrimination (same vs. different comparison task), and one pertaining to categorization (assignment under a common

lexical category). We look first at the psychology of discrimination, and at how the notion of just noticeable difference introduced by Fechner can be related to the tolerance principle, namely the idea that some differences can be so small as to make no difference in terms of discrimination. We then look at the psychology of categorization proper, and review how small differences in terms of similarity to a prototype affect decisions of membership to a category.

In the remaining sections we survey various lines of experimental work based on transition series between distinct prototypes. Such series, omnipresent in several domains of experimental psychology, involve so-called morphs, namely gradual alterations of a prototype connecting it to another prototype. Section 3 looks at two paradigms involving unordered presentations of stimuli drawn from such morphing series: the first concerns studies on *categorical perception*, the second concerns studies of the effect of *simultaneous presentation* of stimuli on categorization. Finally, section 4 surveys work on *dynamic sorites*, that is on ordered transitions between prototypes.

To highlight the importance of such transition series in psychology, we deliberately reproduce several examples of stimuli in this chapter. One message of this chapter is that the manner in which such stimuli are presented (whether isolated, in pairs, in random order, or in a specific order) is essential to the way in which similarity between stimuli influences their assignment to a common category.

1. Slippery slopes and the psychology of reasoning

Traditional definitions of a sorites argument distinguish a narrow sense and a broad sense of the term. In the entry "Sorites" of Peirce and Baldwin's *Dictionary of Philosophy and Psychology*, two senses are distinguished in that way (Peirce & Baldwin, 1902). On a specific and marked sense, it is a particular fallacy, namely the sophism of the heap of wheat usually credited to Eubulides of Megara. On a generic and neutral sense, a sorites is merely a "chain of syllogisms".

The two meanings are obviously related, because the sophism of the heap can be presented as such a chain of syllogisms. Le Chevalier de Jaucourt, in the earlier *Encyclopédie* of Diderot and D'Alembert, writes about the argument of the heap: "that argument is composed of several propositions, differing little from one another, and chained in such a way that, after beginning with a manifest and incontrovertible truth, one moves, little by little, to an obviously false conclusion" (cited in Cayrol, 2016). However, not all chains of arguments need be faulty according to the broad definition of a sorites. Le Chevalier de Jaucourt, in the same entry, mentions a number of precautions that one may take in order for a chain of arguments, that is a sorites in the generic sense, to preserve the truth of its first premise down to its final conclusion.[1]

[1]He writes:

> "To avoid surprise, one needs to ensure that everything that is said of the attribute be also said of the subject. That there be no ambiguity in the terms, nor in the propositions. That one insert no negative propositions among affirmative ones. That the proposition that immediately precedes the conclusion not be negative, unless the conclusion might also be negative. That the link and gradation that must be between the propositions be

One area of particular interest in relation to the previous definitions concerns the psychology of reasoning, and the analysis of so-called slippery slope arguments (or SSAs for short). Different forms of SSAs have been distinguished in the literature, two of which are sometimes called *horrible result SSA*, and *arbitrary result SSA* (see Williams, 1985; Lode, 1999; Volokh, 2003), which we may present as follow:

(1) If A_0 then A_1; if A_1 then A_2; ...; if A_{n-1} then A_n; but A_n is bad; therefore A_0 is bad. (horrible result SSA)

(2) If A_0 then A_1; if A_1 then A_2; ...; if A_{n-1} then A_n; therefore there is no $i < n$ for which it is rational to have that A_i and not A_{i+1}. (arbitrary result SSA)

Both arguments forms are soritical in that they rely on the existence of a "series of gradual intervening steps" (Hahn & Oaksford, 2006) between the antecedent and the consequent of each conditional. However, two arguments are not interchangeable. They do not yield identical conclusions and they are used for different purposes. The first one is used to prescribe drawing a line (at the origin of the sorites sequence), whereas the second type indicates that a line cannot be drawn and bolsters scepticism.

An early illustration of a slippery slope argument can be found in Bossuet's treatise of logic addressed to the Dauphin (Bossuet, 1677, see Cayrol, 2016, to whom we are indebted). Bossuet uses the following example to illustrate the definition of a sorites as a "heap of propositions":

> "Whoever authorizes violent enterprises ruins justice; whoever ruins justice breaks the link that unites the citizens; whoever breaks the society link generates divisions within a state; whoever generates divisions within a state exposes it to an obvious danger; therefore, whoever authorizes violent enterprises exposes the State to an obvious danger".

Bossuet's example is best cast in the form of a horrible result SSA (by adding the premise "but exposing the State to an obvious danger is bad", and then by adding a further conclusion of the form: "therefore authorizing violent enterprises is bad"). A case that more easily lends itself to either type is the following:

(3) a. If abortion may be legal at 1 week of pregnancy, then it may as well be legal at 2 weeks; but if it may be legal at 2 weeks, it may as well be legal at 3 weeks; ...; but if it may be legal at 31 weeks, then it may be legal at 32 weeks; but making abortion legal at 32 weeks is bad. Therefore, that abortion may be legal at 1 week of pregnancy is bad.

b. If abortion may be legal at 1 week of pregnancy, then it may as well be legal at 2 weeks; but if it may be legal at 2 weeks, it may as well be legal at 3 weeks; ...; but if it may be legal at 31 weeks, then it may be legal at 32 weeks; therefore,

---

right. Finally, that there be in the sorites no particular proposition, except maybe for the first. Such are, in brief, the wise rules that Facciolati has detailed in a discourse on insoluble arguments; one can consult it".

> there is no week such that it is rational that abortion may be legal that week and not the next.

SSAs are generally presented as fallacies or sophisms. The general thought that SSAs, like soritical arguments more generally, are incorrect arguments, is epitomized in the following passage from (Blackburn, 2002):

> "Slippery slope reasoning needs to be resisted, not just here but everywhere. It is exemplified in the paradox of the bald man, known as the Sorites paradox. (...) Consider the imposition of a speed limit, say 30 miles per hour, and make it the law. We do not really believe that 29 miles per hour is always safe, and 31 is always not. But we would not listen to someone saying, 'There is no principled place to draw a line, so we can't have a limit'. Nor would we listen to Sorites reasoning forcing the limit forever upwards, or forever downwards to zero. So, if we think the abortion issue does need moralizing and politicizing, nothing stops us from fixing a particular term of pregnancy beyond which abortion is generally prohibited."

In this passage, Blackburn argues primarily against arbitrary result SSAs, suggesting that they are always fallacies. On the other hand, as pointed out by Lode (1999) and Hahn and Oaksford (2006), horrible result SSAs need not be fallacies. They may be seen as instances of a broader class of "empirical" or "rational grounds" SSAs, which can be used to rationally argue in favor of drawing the line at the origin. Bossuet's example, clearly, is not intended as a fallacious argument, but rather as a compelling argument leading from sound conditional premises to a sound conditional conclusion. In principle therefore, SSAs ought to fall under the generic-neutral definition of a sorites: like other chains of propositions, they can have sound instances and unsound instances, although the unsound instances will create more trouble, and will generally be seen as more emblematic of the notion.

Hahn and Oaksford (2006)'s main point is that the acceptability of "empirical" or "rational grounds" horrible result SSAs varies as a function of the strength of the probabilistic connection between the consequent and the antecedent in the conditional, and as a function of the negative utility of the outcome (how bad or "horrible" the outcome is supposed to be). This was experimentally confirmed by Corner, Hahn, and Oaksford (2011) for horrible result SSAs of length 1, by having participants rate the acceptability of argument strength as a function of the utility and probability of outcomes.[2]

The results of Corner et al. (2011) can be brought to bear on the discussion of arbitrary result SSAs more broadly. That is, the consideration of the probability of conditional premises can serve to determine where the line should be drawn. Consider legal dispositions on abortion. Depending on the place, the line is effectively drawn between antecedent and consequent at values (in terms of age of embryo) for which, despite equal steps in weeks,

[2]Corner et al. (2011) do not use explicit conditional sentences to test those predictions, but they use related constructions (e.g. "We should oppose the legalisation of euthanasia in the UK, as it will lead to an increase in the number of instances of medical murder"').

the developmental differences appear larger. To put it otherwise, the conditional statement "if you allow abortion at 1 week, you may allow it also at 2 weeks" appears to have higher validity than the conditional statement "if you allow abortion at 5 weeks, you may allow it also at 6 weeks", if it is felt that the strength of the association between 5 and 6 is smaller than the strength of the association between 1 and 2, and possibly even weaker than at other places. An example is Ohio's "hearbeat bill", drawing the line between 5 and 6 weeks on account of the emergence of a noticeable heartbeat in embryos during that step.

More generally, the strength of an SSA ought to depend on the inductive strength of each conditional premise, and on the length of the chain, since intuitively the longer the chain, the weaker the probabilistic connection between the last consequent A_n and the first antecedent A_0.[3]

On Hahn and Oaksford's account, SSAs thus fall under a broader Bayesian account of inference and argumentation. Such an account can explain the sensitivity of other kinds of arguments to context. From our perspective, there are two virtues of the account of slippery slopes outlined by Hahn and colleagues using the notion of conditional probability: one is the fact that such an account avoids rejecting soritical reasoning as always flawed. The other is that it highlights a connection between soritical reasoning and probabilistic reasoning, which we will see to be of importance in other areas of the psychology of the sorites paradox.[4]

2. Discrimination and categorization

Viewed abstractly, the sorites paradox can be presented as a puzzle concerning the impact of similarity on judgment. The main premise of a sorites argument says that if two objects are sufficiently similar, then they will be judged alike. To say that they will be judged alike can mean two different things: that similar objects will be treated alike in terms of discrimination (same vs. different recognition), or that similar objects will be treated alike in terms of categorization (assignment under a lexical category).[5] In this section we consider how both principles are approached in psychophysics. The focus in this section is mostly foundational, and concerned with the centrality of probabilistic modeling for an adequate representation of both discrimination and categorization.

2.1. **Similarity in discrimination.**

Whether and in what sense two very similar objects will be judged alike is in a way part of the initial project of psychophysics (Fechner, 1860).

[3]In logic, this feature may be characterized in at least two ways: either in terms of a conditional connective showing failures of transitivity; or in terms of a consequence relation failing transitivity. On the former kind of approach, see for instance Adams (1998); on the latter see for example Cobreros, Egré, Ripley, and van Rooij (2012).

[4]See in particular Lassiter and Goodman (2015) for a more recent Bayesian account of soritical reasoning with vague adjectives. Their account compares various ways of formalizing the main conditional premise of the sorites in probabilistic terms. On their account too the size of the step is crucial in determining the probability of the main conditional premise of the sorites, which is typically less than 1.

[5]Terminology varies in psychology: studies on categorical perception oppose discrimination tasks and identification tasks. In this paper we preferably use the term "categorization" instead of "identification", to refer to the assignment of an item under a higher-type category.

Fechner was interested in measuring the effect of variations of physical magnitudes on perceived magnitudes. For example, he was interested in *the extent* to which a weight physically heavier than another would be perceived as heavier. Importantly, Fechner did not concern himself exclusively with small variations of physical magnitude on perceived magnitude, but with the general problem of the relation between the two kinds of magnitude. As a limiting case, however, Fechner was interested in the problem of the minimal difference in physical magnitude that it would take for a difference in sensation to be perceived (what he called a Just Noticeable Difference or JND). On the assumption that such a difference exists, we may state the relation between physical difference and perceived similarity as follows:

$$(4) \qquad |w(x) - w(y)| \leq \varepsilon \rightarrow x \sim_W y.$$

This says that if the difference in weight between x and y is less than some positive value ε, then x and y will be qualitatively perceived as having the same weight. Whether there is a positive value ε of physical difference along some relevant dimension, such that *absolutely* no difference will be perceived is a difficult problem that quickly aroused discussion among Fechner's contemporaries. This problem obviously bears a connection with what Wright (1976) has called the *tolerance principle*, the idea that there might be some positive degree of change of some property "insufficient ever" to make a judgmental difference.

The way this problem was solved, already by Fechner, is by appeal to statistical methods. The observation of psychophysicists is that even when two stimuli are successfully discriminated along some sensory continuum on one or several occasions, there remains a probability of confusing them. That probability would materialize in terms of the number of failures to discriminate the stimuli over sufficiently many trials. Conversely, even when two stimuli are very hard to discriminate, one may find evidence that they are not perceived as entirely alike by running sufficiently many trials and by looking at the proportion of success at discrimination.[6] Because of that, the common wisdom in contemporary psychophysics is to think of what counts as a "just noticeable difference" as being relative to a probabilistic threshold, whose choice is not unique but is set conventionally. Luce (1959, 34) presents the idea of JND as follows:

> "The essential idea is to pick a probability cutoff π, $\frac{1}{2} < \pi < 1$, and to say that alternatives discriminated more than 100π per cent of the time are more than one JND apart; those discriminated less often are one JND or less apart. (...) That is to say, it is meaningless to speak of JNDs without specifying the probability cutoff that was used to define them – a point unfortunately all too often ignored in the experimental literature."

On that approach, $x \sim_W y$ is thus definable in terms of the probability of confusing x and y relative to a statistical threshold, and the value of the constant ε is in fact relative to that threshold. That is, $x \sim_W y$ iff $1 - \pi \leq Pr(x, y) \leq \pi$, where $Pr(x, y)$ is the probability

[6]See Borel (1950) and Hardin (1988) and Raffman (2011) for discussions of that issue.

of selecting x within the set $\{x, y\}$ (see Luce, Definition 3).[7] For illustration, suppose that $\pi = 0.79$, then x will be declared noticeable from y if x is selected more than 79% of the time, or if y is selected more than 79% of the time.

In psychophysics various paradigms exist for the measurement of that probability, such as the 2-alternative forced choice task (2AFC), in which for example two color patches x and y are presented on a screen, and in which participants must decide whether a third patch z is identical to x or y. The patch z is always one of x or y: typically, when x and y are physically similar, it will be hard to have accurate matching judgments, and the proportion of correct answers will allow the experimenter to decide whether x and y are more or less than 1 JND apart. Note that on that approach of the definition of a JND, the highest degree of confusion is when the confusion probability between x and y is close to 0.5, meaning the capacity to discriminate between x and y is at chance level.

According to Dzhafarov and Dzhafarov (2012) a *comparative* sorites sequence is a sequence of stimuli $(x_1, ..., x_n)$, such that adjacent stimuli in the sequence are pairwise indiscriminable in the sense of being in the relation $x_i \sim_W x_{i+1}$, but such that x_1 and x_n are not in that relation. Luce (1956) points out that we find an abundance of such comparative sequences in which one is indifferent between adjacent members of the sequence, but not indifferent between more distant members (Luce's topical example involves a series of 401 cups of coffee with increasing amounts of sugar, such that we can't distinguish the sweetness of adjacent cups of coffee, but we can definitely taste the difference between the cups with the smallest and largest amounts of sugar). This implies that indifferences are not transitive, a point central to Luce's account of indifference, and captured by his approach via probability cutoffs: it is easy to find triples x, y, z of stimuli such that $Pr(x, y)$ and $Pr(y, z)$ are both below π but $Pr(x, z)$ is not. This means that two or more members of a sequence can be less than one JND apart, while the ends of the sequence are more than one JND apart.[8]

2.2. **Similarity in categorization.** The conditional (4) states that if the physical difference between two objects is small enough, then the perceived difference between them will be small. (4) should be compared with the standard premise of the sorites paradox, which may be put as follows:

$$(5) \qquad x \sim_W y \rightarrow (P_W x \leftrightarrow P_W y).$$

Whereas (4) says that a small *physical difference* produces no difference in *discrimination*, (5) states that a small difference in discrimination makes *no* difference in the assignment under an abstract category. For example, if the difference in perceived weight (W) between

[7] Luce also defines the relation $xL_W y$ as $P(x, y) > \pi$. Intuitively, it means "at least one π-jnd larger", and $x \sim_W y$ thus means "no more than one π-jnd apart", the latter being a reflexive, symmetric, but typically nontransitive relation.

[8] Luce (1956) contains an algebraic account of the notion of intransitive preference, and of its relation to the corresponding of preference, in what is known as the theory of semi-order relations. See van Rooij (2011) for a presentation of Luce's account.

two objects is small enough, then both should be declared "heavy" (P_W), or both should be declared "not heavy".

Dzhafarov and Dzhafarov (2012) relate the latter premise to what they call *categorical* sorites sequences: a categorical sorites sequence is sequence of stimuli $(x_1, ..., x_n)$ such that adjacent members are in the $\sim_W$ relation, and yet such that x_1 satisfies P_W but x_n does not. Such sequences undeniably exist, but they cannot exist consistently with the admission of principle (5), assuming the logic and the conditional to be classical.[9] The existence of such sequences is in fact the reason why the literal interpretation of principle (5) is rejected by so many accounts of the sorites (viz. Borel, 1907, 1950; Williamson, 1994; Raffman, 2014; Dzhafarov & Dzhafarov, 2012).

What is the situation in psychophysics? The dominant view is that in the same way in which the relation $x \sim_W y$ expresses a probabilistic dependence in the discrimination between x and y, the biconditional in (5) should be weakened to a relation expressing a probabilistic dependence in the identification of x and y under a common category. In other words, (5) should be weakened to:

(6) $x \sim_W y \rightarrow (P_W x \approx P_W y)$, where $P_W x \approx P_W y$ iff the probability of judging $P_W x$ is close to the probability of judging $P_W y$.

In the case of weight, this says that if two objects have almost identical weights, then the probability of categorizing each of them as heavy will be almost identical (see Egré, 2011a). Note that this principle is weaker than (5) because it does not prevent small differences along the dimension of similarity from making a difference along the dimension of categorization. When we combine (6) and (4), we see that if two objects are sufficiently similar to be indiscriminable, then the probabilities of subsuming them under the same category will be relatively close, but that is not to say that the objects will invariably be assigned to the same category (Borel, 1907). One way of summarizing this is to say that psychophysical models of categorization do not endorse the tolerance principle (5), but they nevertheless support a probabilistic version of what Smith (2008) has called the *closeness* principle. Smith's closeness principle says that if two items x and y are very close in terms of their P-relevant properties, then the truth values attached to Px and Py should be close. Assuming only two truth values are available to begin with, a different way of cashing out that idea is in terms of probabilities (instead of degrees of truth).[10]

The upshot is that items that are highly similar *may* but *need not* be categorized in the same way in all circumstances. When imagining two very similar color patches that

[9] We have deliberately modified the exact definition of a categorical sorites sequence given by Dhzafarov and Dhzfarov. On their definition, it directly follows that such sequences cannot exist. We prefer to say that sequences whose adjacent members are hard to tell apart exist, but to highlight that the tolerance principle, classically interpreted, is problematic in combination with such sequences.

[10] See Egré (2011b) for an interpretation of closeness along those lines, and Lassiter and Goodman (2015) on various ways of articulating the tolerance principle probabilistically in a Bayesian setting. This is consistent with the idea of representing the tolerance principle more qualitatively than quantitatively. See Cobreros et al. (2012) for a qualitative version of the tolerance principle, and Egré (2011a) and Egré (2016) for some bridges between qualitative and probabilistic representations.

would be indiscriminable, one is strongly pulled to the view that either both of them will necessarily be perceived and categorized as red, or that neither of them will be. Raffman (1994) tempers that intuition by stressing that discrimination and categorization obey distinct constraints:

> "If we are mystified by our ability to draw categorial distinctions between patches we can't tell apart, that is partly because we are setting things, as it were, from the discriminator's point of view. The discriminator has no memory to speak of, and certainly no memory of the sort required for categorial distinctions. Hence he fails to notice the progressive change in the appearance of the patches as he moves along the series. The categorizer, on the other hand, has a rather good memory (...). We might capture the idea by envisioning the category as a mental elastic band anchored at one end to the stored prototype."

Raffman's view here is faithful to the idea that categorization of an item is a function of the similarity of that item to a prototype stored in memory. Discrimination, on the other hand, is fundamentally a local process of comparison.

2.3. **Threshold models.** Various probabilistic models of categorization exist in psychology that satisfy the constraint stated under (6). In this section we consider two examples belonging to the family of *threshold* models (Hampton, 2007; Verheyen, Hampton, & Storms, 2010; Verheyen, Dewil, & Egré, 2017; Borel, 1907; Egré, 2016), introduced explicitly to handle vague predicates. Such models view the subsumption of a stimulus under a category as relative to an inner threshold value lying on a continuum. Category membership is fixed by whether the stimulus is perceived as above or below the relevant threshold along the relevant dimension, but interstimulus differences are handled probabilistically.

The first example we consider is a pioneering model outlined in Borel (1907) in relation to the sorites paradox. Borel's model is a Gaussian model of categorization as a function of imperfect discrimination. When discrimination is perfect, observation will be exempt from noise, and category membership will be represented by a step function (Figure 1, left): any item above the threshold will be a category member, and any item below a nonmember. When discrimination is imperfect, on the other hand, category membership is represented by a smooth function as a result of noise: the typical form of such a function is a sigmoid function derivable as a cumulative normal distribution function (see Figure 1, right) with the inflection point centered on the threshold, and whose slope at that point depends on the amount of noise in discrimination.[11] As Figure 1 exemplifies, small variations in the physical properties of the stimulus (such as height in cm) are matched by small variations in the probability of categorizing an item under a given category (such as "tall"), in agreement with principle (6). However, the relation between physical similarity and categorization is not linear, since identical differences along the physical axis can be matched by more or less difference in the probability of categorizing an item under a given label.

[11] See Egré and Barberousse (2014) and Egré (2016) for details on the derivation of the function from first principles.

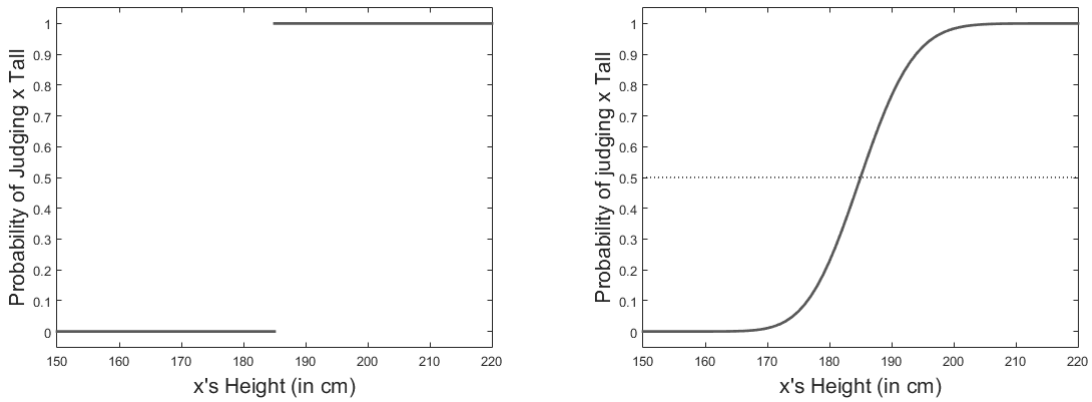

FIGURE 1. Hypothetical categorization curves for "Tall" as a function of physical height. On the left, a step function, with a threshold at 185cm; on the right, a smooth function, based on the cumulative normal distribution relative to the same threshold.

The second example we consider is Hampton (2007)'s threshold model of the relation between membership and typicality. In Hampton's approach, as in Raffman's picture of categorization, similarity is relative to a prototype for the category. The decision whether to categorize an item as P or not P is therefore a function of the similarity of that item to the representation of the prototype in memory. When the similarity exceeds a certain inner threshold, the model predicts a verdict of membership. As shown by Hampton (1998), actual data collected by McCloskey and Glucksberg (1978) concerning the relation between binary membership judgments and typicality ratings are adequately fit by a cumulative normal distribution. McCloskey and Glucksberg collected binary membership judgments of 30 participants and typicality judgments from a distinct group of 24 participants about 492 items taken from 18 distinct categories. Hampton (1998) represented average judgments of membership across participants as a function of the typicality ratings obtained for the other group (Figure 2, left). What they found is that items most typical of a category (viz. a car relative to the category vehicle, or a diamond relative to the category of precious stones) and items entirely atypical for a category (such as shoes for vehicles, or granite for precious stones) have a very high probability of being respectively included in or excluded from the category. For items of intermediate typicality (say a parachute for the category vehicle, or zircon relative to precious stones), the degree of membership is itself intermediate.[12]

Figure 2 shows the distribution of average membership judgments as a function of typicality for the separate categories "Vehicle" and "Precious Stone". In each case, membership is indeed a sigmoid function (represented by the best fitting logistic function). As those

[12]See Verheyen et al. (2010) for a replication and probabilistic account of the McCloskey and Glucksberg findings, and Verheyen, De Deyne, Dry, and Storms (2011) for an extension that includes (dis)similarity to prototypes of two competing categories.

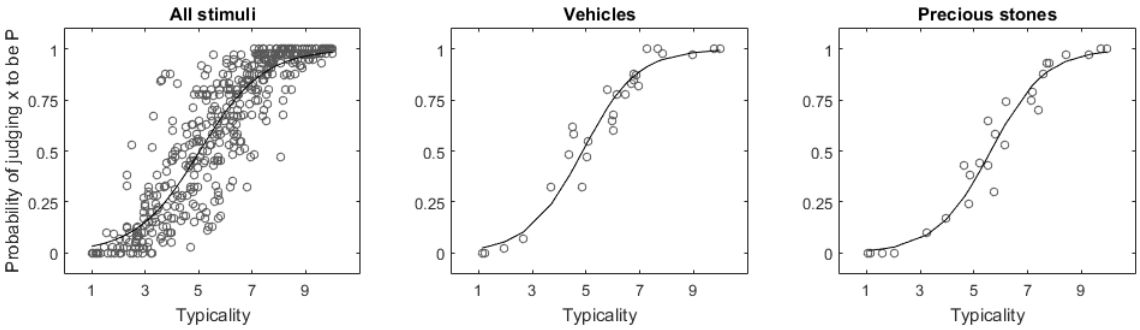

FIGURE 2. Scatterplots of the percentage of participants in McCloskey and Glucksberg (1978) providing a positive categorization versus average item typicality for all stimuli (left), the category of vehicles (middle), and the category of precious stones (right) along with the best fitting logistic function.

data confirm, small differences in typicality again make small but non-null differences in membership judgments. This pattern, importantly, is not just operative at the group level. When binary membership judgments are collected within-subject, and the same item is presented multiple times to a participant, the average degree of membership of each item can be calculated in the same way, and individual psychometric functions present the same characteristic shape (see for example Egré, De Gardelle, & Ripley, 2013, for individual data pertaining to color terms).

2.4. **Typicality and family resemblance.** To conclude this section, we may highlight further aspects of the influence of the sorites paradox concerning the centrality of the notion of prototype in categorization. One point of contact worth mentioning between philosophy and psychology around the sorites can be found in Wittgenstein (1953)'s remarks in §65 to 78 of the *Philosophical Investigations* on the limits of criterial definitions, and in the subsequent work of Rosch and Mervis on the notion of family resemblance (Rosch & Mervis, 1975).

Wittgenstein in this passage argues for the impossibility of defining a concept such as "game" by the possession of necessary and sufficient criteria of application. Instead, Wittgenstein argues, card games, ball games, and other kinds of games form a family in which "we see a complicated network of similarities overlapping and criss-crossing" (§66). Various games can therefore be ordered depending on how many features they have that overlap. As Rosch and Mervis (1975, p. 575) summarize,

> "a family resemblance consists of a set of items of the form AB, BC, CD, DE. That is, each item has at least one, and probably several, elements in common with one or more other items, but no or few elements are in common to all items".

Basically, therefore, a family resemblance is a set of items with soritical structure, in the sense that each such set can be ordered with adjacent items having one or more features

in common, but such that the first and last item in the sequence may have no features in common. What Wittgenstein's remarks suggest is that the limits of a concept such as "game" could be hard to find precisely in virtue of that soritical structure, namely in virtue of the fact that items pairwise are highly similar. Rosch and Mervis appear to agree with that, but what they point out, and what McCloskey and Glucksberg's data confirm, is that in cases in which a concept appears to have no clear boundary in relation to that family resemblance structure, the degree as well as the ease to which membership is decided is a function of the centrality of the item in the category. There is more, in other words, to the notion of family resemblance than the existence of local similarities: there is the idea that among a set of items with a family resemblance, some items are more stable than others relative to the assignment to that common category. Rosch and Mervis operationalize this in terms of what they call a *family resemblance score*. For example, "chair, sofa, table,..., lamp, telephone" all have various features that may justify assigning them to the category "furniture", but in that list the more typical items are in fact also those that share the most features with other items in the family (i.e. those that have a higher family resemblance score).

On the one hand, therefore, local similarities in a set of items with family resemblance structure might explain why it is so difficult to draw nonarbitrary boundaries for the corresponding category. On the other hand, this difficulty seems helped, in practice, by the fact that some items are much more central than others in the set, and can serve as reference points to decide category membership. Rosch and Mervis's notion of family resemblance is particularly relevant for nominal or multidimensional categories, for which several features can be distinguished, even if the notion of typicality is applicable to a wider range of sorites-susceptible cases, including adjectival categories (see, for instance, Verheyen & Egré, 2017). Another aspect worth stressing is the fact that in their studies on nominal categories, items are all in general pairwise discriminable (viz. an apple vs. an avocado relative to the category "fruit") unlike stimuli generally examined in relation to adjectival categories (viz. two adjacent shades in relation to the category "yellow", or two close heights relative to "tall"), but categorization can nevertheless give rise to inconsistent verdicts both between- and within-subjects (considering both an apple and an avocado to be fruit on one occasion, but denying that avocado is fruit on another; see McCloskey & Glucksberg, 1978, and Verheyen et al., 2010).

3. Unordered transitions between categories

In this section we examine the relation between discrimination and categorization in experimental paradigms in which intermediate stimuli drawn from gradual transition series between two prototypes are presented in a random order. We highlight that aspect since randomness is generally a way of blocking specific order effects, which we will discuss in the next section. We look at two sets of phenomena that suggest that the relation between discrimination and categorization is more complex than what is assumed in the main premise of the sorites paradox. We first discuss the phenomenon of categorical perception.

We then refine the discussion in the previous section by reporting on the effect of presenting pairs of more or less distant stimuli on the ascription under a common category.

3.1. **Categorical perception.** The main premise of a (categorical) sorites (5) asserts that if two color patches, say, are indiscriminable, then they will be assigned the same abstract category. As Raffman's earlier quote warns us, that view presupposes that discrimination is mostly what drives categorization. One phenomenon of particular interest in that regard is categorical perception. The phenomenon concerns the converse influence that categorization (assignment under a linguistic label) appears to exert on discrimination. Harnad (1987) describes it as follows:

> "The effect is best described as a qualitative difference in how similar things look depending on whether or not they are co-classified in the same category".

More specifically, categorical perception involves two components. The first is the fact that, over a variety of sensory continua, subjects are able to maintain a reliable category boundary within each continuum. The second is the observation that for stimuli that are *physically* equally spaced along the relevant continuum, discrimination across categories appears to increase over discrimination within each category ("category boundary" effect). The effect was first observed in the perception of phonemes (see Repp, 1984, for a review), but has been documented since in a number of other domains, in particular in the perception of colors and facial expressions (Calder, Young, Perrett, Etcoff, & Rowland, 1996).

The first experiments in the case of phonemes concerned the perception of syllables with distinct onset consonants. Liberman et al. were able to construct a series of syllables interpolated between syllables unambiguously identified as /be/ or /ge/.They tested both discrimination performances of participants and categorization performances. They measured discrimination using an ABX paradigm: participants were presented with triads of sequential phonemes, first A, then B, and then a third phoneme X which was one of A and B, and which they had to judge as being identical either to A or to B.[13] In the second part of the experiment, participants had to identify each phoneme along the continuum as either /be/ or /ge/. Both in the discrimination task and in the categorization task, stimuli were presented in random order along the continuum.

The notion of category boundary in the categorization task is defined as the position of the point of subjective equality (50% of trials categorized as /be/) along the continuum. Discrimination performance is measured as the percentage of correct responses over consecutive pairs along the stimulus set. The main observation was an increase in discrimination performance for stimulus pairs that straddle the category boundary, compared to stimulus pairs on either side of the boundary, despite equal physical distance. Figure 3 presents data collected by Calder et al. (1996) based on a stimulus set involving 11 pictures, with two prototypes of either a sad or a happy face (see Figure 3) with 9 morphs interpolated

[13]The ABX paradigm can therefore be viewed as a sequential 2AFC paradigm (see section 2.1).

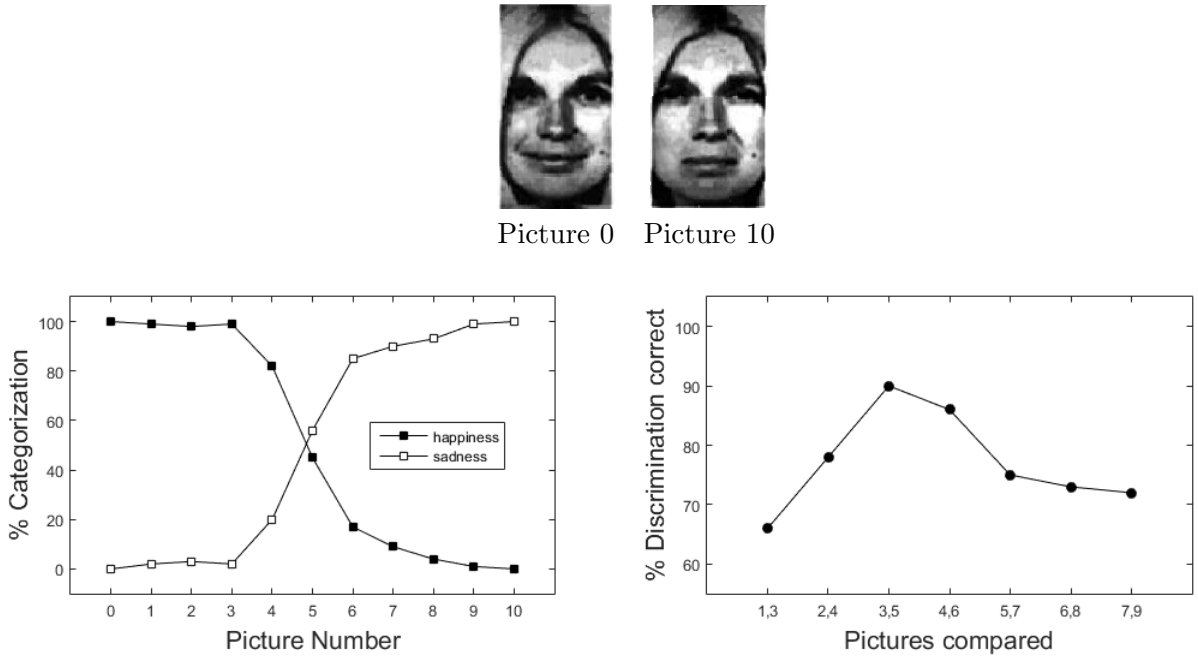

FIGURE 3. Categorization and discrimination data for pictures of faces ranging from Happy (0) to Sad (10), based on Calder et al. (1996)'s data, Fig. 2. Discrimination shows a peak toward the point of subjective equality for categorization.

between them. As the figure shows, discrimination increases over pairs nearer the point of subjective equality for categorization, and decreases on either side of that threshold.

The phenomenon of categorical perception raises several questions which we can only briefly mention in this chapter. The first is how much the phenomenon is sensitive to the discrimination task used. Repp (1984) reports variability depending on the structure of the task (considering various alternatives to the ABX task), but concludes that the phenomenon is overall robust. The second is whether the same phenomenon occurs for all kinds of sensory continua and for all subjects. The answer to that question appears to be negative, since discrimination does not always show a peak depending on the stimulus and the perceiver (Repp, 1984). A third question is whether it is adequate to describe categorical perception as an effect of categorization on discrimination, rather than the other way around.[14] It is widely considered to be the case, for classical psychophysics assumes discrimination to be monotonically decreasing for stimuli with a fixed physical difference (on account of Weber's law, which states that perceived changes are proportional to the initial stimulus), but the question is widely debated (Macmillan, 1987). A related question one may ask is whether categorical perception might be an artifact of the assumption

[14]See also Quine (1970) and Verheyen and Storms (2011).

that stimuli are one-dimensional (Macmillan, 1987). As Figure 3 shows, discrimination performance for facial expressions can reach 90% at the category boundary, which might exceed the relevant threshold used to define a JND. If that is so, however, one may think of a different way of constructing the stimulus set, where the steps along the continuum are not just physically equal, but perceptually equal (each leaving performance at the same level). Doing so, however, would involve densifying the region of intermediate stimuli along the salient physical dimension in order to leave discrimination approximately constant.

While the phenomenon invites much more discussion than we can go into here, it is important to acknowledge that the findings related to categorical perception have spurred the development of probabilistic accounts of categorization (see for instance Macmillan, 1987 for a survey). A more substantive perspective is that the assignment under a category may not be solely a function of discrimination relative to a threshold, but may actually obey separate constraints.

3.2. **Categorization of pairs of stimuli.** Classic studies of categorical perception generally investigate the categorization of stimuli sequentially. What happens, however, when a stimulus is presented simultaneously with a similar one? That question is of particular interest, for some philosophers of vagueness have argued that in a context in which two items are "saliently similar", they should be categorized alike. Fara (2000) calls this principle the *similarity constraint*. The similarity constraint is a weakening of the standard tolerance principle, for it restricts it to the simultaneous presentation of two objects. We may represent it as follows:

(7) If $x \sim_W y$ *and* x, y are presented side by side, then *in the context of that presentation* $P_W x \leftrightarrow P_W y$.

This says that two objects that are hard to tell apart, *when presented side by side*, should be assigned the same category in that specific context.[15] Prima facie the effect of a joint presentation should be to enhance similarity, and therefore to encourage the assignment of two adjacent stimuli in a sorites sequence under a common category, but one may wonder if that is really so, in particular for borderline cases showing less stability over successive trials.

We do not know of any study that addresses exactly that question, but two papers by Hampton and colleagues are of relevance to that issue (Hampton et al., 2005; Hampton, Estes, Botbol, & Jaunbocus, 2012). Hampton and associates investigated the influence of the presentation of a stimulus of varying similarity on the categorization of a borderline stimulus, both within and across the category boundary. In their leading experiment, Hampton et al. first presented participants with 9 color patches spanning the region between a blue prototype and a purple prototype, in a random order. For each participant, they determined the borderline patch b closest to the participant's point of subjective equality, taken to represent the underlying participant's category boundary. In a second

[15] See (Egré, 2011a, 86) for a more explicit formalization of this constraint. The effect of the relativization is that in the context $\{a, b\}$, a and b may both be judged P_W, but in the context $\{b, c\}$, b and c may both be judged $\neg P_W$.

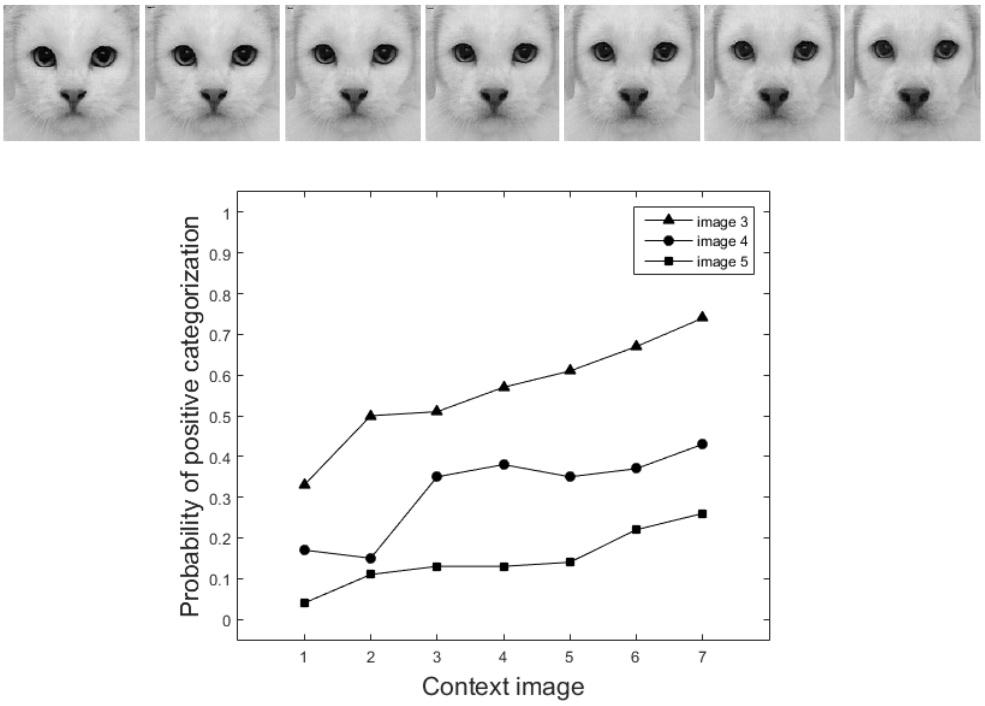

FIGURE 4. Hampton, Estes, and Simmons (2005)'s cat-dog morphing series and mean probabilities of categorization of target image as *cat* as a function of context image (based on Hampton et al., 2005, Figure 6).

phase, participants were shown pairs drawn from 7 out of those 9 patches (the prototypes were excluded), at varying distance from each other in the series. Participants had to decide whether only the right patch, only the left, both, or neither was blue (respectively purple).

The responses of main interest concerned the categorization of the borderline patch b relative to context patches $b{+}1$, $b{+}2$ and $b{+}3$ (toward the purple end), and similarly relative to context patches $b-1$, $b-2$ and $b-3$ (toward the blue end). When the context patch was more than one step away, what was found was a robust contrast effect: the probability of judging b as of the same category as the context hue was decreased in both directions. Only when the context hue was one step away was a tendency toward assimilation observed, and that assimilation effect amplified when participants were given the choice between only "both" and "either" answers (Hampton et al., 2012). The contrast effect was replicated with different choices of colors and also with morphs presenting ambiguous stimuli between a typical dog face and a typical cat face (see Figure 4). Figure 4 presents the mean probabilities of categorizing the middle pictures of that stimulus set as a function of other context pictures. For example, when the target image is compared with the most cat-like picture (1), we see that participants will tend to categorize it less as a cat, whereas when

the target image is compared with the most dog-like picture (7), they tend to categorize it more as a cat.

Regarding the constraint of salient similarity, Hampton et al's data nevertheless suggest that for adjacent items from a series of morphed images presented together, there will be a tendency toward assimilation, namely toward assignment in the same category. This effect is fragile, however, for as soon as the step increases, the context picture will tend to push the borderline stimulus outside the category to which the context picture belongs. We note also that Hampton et al. do not report the mere proportion of "(only) left" and "(only) right" responses when target is picture x_i and context is x_{i-1} or x_{i+1}. It would be interesting to check if that proportion stays approximately constant across the range, or increases toward the category boundary (that is, with participants assenting less to "both" or "neither" judgments than they do toward the prototypes). If it stays constant, this would be a more direct confirmation of the constraint of salient similarity. Otherwise, coherently with categorical perception effects, we may have to conclude that the constraint of salient similarity itself is modulated by how distant the items compared are to the prototype of the category.

4. Ordered transitions between categories

In section 2 we considered the problem of how similarity to a prototype impacts categorization when stimuli drawn from a sorites sequence are presented in no specific order. In the previous section we looked at the more specific problem of how salient similarity can favor assimilation under a common category in borderline cases in similarly unordered conditions. In this section we consider actual experiments done on so-called *dynamic sorites* or *forced-march sorites* (Horgan, 1994), namely ordered transitions between two prototypes by small steps. Ordered transitions of that kind are of particular interest, because they pose the problem of how and when a category switch occurs along a sorites sequence. We first consider dynamic transitions and their effect on discrimination, and then continue with the effect they have on categorization.

4.1. **Slow motions and change detection.**

Some case studies exist in psychology that involve what we may call *dynamic comparative sorites*, namely ordered sequences of stimuli in which the task is for participants to report whether there is a change or not. The task, in those cases, is concerned purely with discrimination.

The best example we can think of is in the field of change detection. An abundant literature concerns the phenomenon of change blindness in the visual domain (Simons & Levin, 1997; O'Regan, Rensink, & Clark, 1999). One influential paradigm involves the comparison between two pictures shown in alternation, with a brief flicker or disruption between them, sometimes with the inclusion of some distractors (see O'Regan, 2001, for a review). A more or less extended portion of the original picture scene is altered between consecutive pictures (an object is added or removed), and often participants are unable to detect the change or they take time and effort to correctly report it.

While flickering pictures appear to bear no relation to the sorites, one class of stimuli used to evidence change blindness is of particular interest in relation to dynamic sorites,

and concerns the phenomenon of blindness to slow changes. A demo visible on O'Regan's website (`http://nivea.psycho.univ-paris5.fr/ECS/sol_Mil_cinepack.avi`) and programmed by Renaud Chabrier is a 44" long animation based on the gradual alteration of a still image depicting a carousel. The animation involves a very progressive alteration of the luminosity and then color of the platform of the carousel. The animation provides a striking example in which the image appears not to change at all from one picture to the next, even a few seconds apart (compare at 0" and 2" in Figure 5). By the end of the animation, one may easily fail to have noticed any change, or fail to accurately report where the change took place. When the movie is rewinded straight to its first picture, however, the contrast in luminosity and color with the last picture is sufficiently marked to be immediately noticeable (see Figure 5, compare between 0" and 44").

FIGURE 5. Three images from Chabrier's slow change movie (at 0", 2", and 44")

Slow changes not only confirm that comparative sorites sequences exist, but their main interest lies also in the fact that it is easy when the rate of change is low enough to not even detect a change between the first and the last image. This indicates that dynamically, the similarity between adjacent pictures can supersede the memory of more distant pictures, thereby maintaining an illusion of identity between the first and the last picture. O'Regan does not report specific measurements based on this paradigm, but such examples may be used to vindicate the idea that short of a vivid memory of the first stimulus seen, or short of the right attentional guidance, there is a soritical adaptation effect in such sequences: the similarity between consecutive pictures can dynamically override the dissimilarity between more distant pictures.

4.2. **Ordered transitions between prototypes.** We now consider some studies of dynamic categorical sorites, namely ordered transitions between stimuli in which the main task is to subsume each stimulus seen under one or more categories. We highlight three studies in that domain, whose results are remarkably consistent with each other, and which evidence a robust order effect in dynamic sorites.

4.2.1. *Three studies based on morphing series.* A first influential study in that area originates in work conducted by Raffman on color categorization, conducted with psychologists Lindsey and Brown (Raffman, 2011, 2014). Raffman and colleagues presented participants with a set of 37 color patches interpolated between a typical blue and a typical green patch. They presented the stimuli in 5 distinct orders: random, from green to blue, from blue to green, and then in two 'reversal' conditions, in which participants, as soon as they switched

category along a given order, were marched backward to the preceding stimuli they had just categorized in opposite order. Participants had the option to categorize each stimulus either as "Blue", "Green", or they could opt out by using a third "?" response.

Raffman reports two main findings from that study. The first is that, in the standard ordered conditions, participants tended to switch category earlier from blue to green, and earlier from green to blue, rather than later, compared to the reverse order and the random order. This phenomenon, sometimes called *negative hysteresis*, or *enhanced contrast* between categories, was found earlier by Kalmus (1979) and has been replicated since in Egré et al. (2013) and Stöttinger, Sepahvand, Danckert, and Anderson (2016) (see below). The second main finding is that participants in the reversal condition continued to apply the category to which they had just switched to the preceding patch. For example, a participant who categorized stimulus 18 as blue and stimulus 19 as green would continue to categorize stimulus 18 as green when marched backwards after the switching. To Raffman and colleagues, this phenomenon is indicative of a local form of (positive) hysteresis: participants carry on with the category they just switched to without seeing a discrepancy with their earlier categorization. In line with the earlier analyses of Raffman, this indicates that borderline stimuli in a sorites sequence can receive inconsistent verdicts without contradiction, meaning that the same items can be judged P and not P but on separate occasions (see also section 2.4 on within-subject inconsistency).

In a related study, Egré et al. (2013) ran a similar task, except that they did not include the reversal conditions, and they included fewer items within each sequence (15 items). Instead, they measured participants' responses along two color sets (Green-Blue and Yellow-Orange, see Figure 6), and they ran two different tasks with different groups of participants: a perceptual matching task in which participants were shown triads of color patches and had to decide whether the target was more similar to either of the two typical end shades of the color set to which it belonged, and a linguistic categorization task in which those end shades were replaced by the names of the categories of which they were typical ("Yellow" vs. "Orange", "Blue" vs. "Green"). What they found was a striking contrast in the results of the two tasks. In the perceptual task, participants showed no order effect: that is, the participants' points of subjective equality (point of 50% application of a given category) was situated at about the same location along the stimulus set, irrespective of the order in which the stimuli were presented. In the linguistic task, by contrast, the participants showed the same order effect reported by Kalmus and Raffman, namely negative hysteresis: participants switched category earlier along the continuum when going from Green to Blue (Orange to Yellow), and similarly when going from Blue to Green (Yellow to Orange). Figure 6 shows Egré et al.'s data concerning the average selection of the category "Orange" depending on the order in which the 15 shades were presented. Irrespective of the order in which the stimuli were presented, the psychometric curves show the usual sigmoid shape, but in the linguistic task we see that the point of subjective equality is shifted to an earlier position along the stimulus set when going from Yellow to Orange, as opposed to from Orange to Yellow. The point of subjective equality in the random condition is located between those of the ordered conditions. In the perceptual

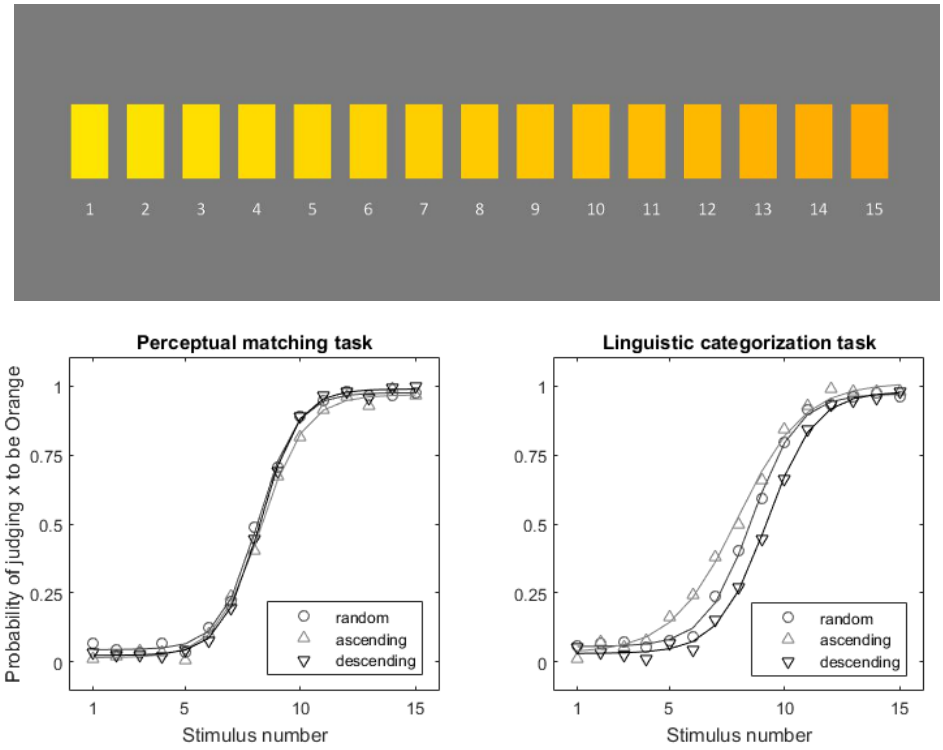

FIGURE 6. Average percentage of "Orange" responses for the 15 color stimuli from Egré et al. (2013)'s Yellow-Orange color set (Top), in the random, ascending, and descending order conditions of the perceptual matching task (Bottom left) and the linguistic categorization task (Bottom right), along with the best fitting logistic functions.

matching task, on the other hand, the psychometric curves basically overlap for all three orders.

More recently, Stöttinger et al. (2016) have designed a class of 40 picture sets each consisting of 15 pictures gradually morphing the silhouette of an object into the silhouette of another object (see Figure 7 for an example). The series were divided into 4 sets, either morphing an animate object to an animate object (such as a cat into a rabbit), an inanimate to an inanimate (such as a pear to a violin), an animate to an inanimate, or the other way around (viz. a bodybuilder to a pair of scissors, or conversely).[16] This time the authors used a naming task: participants had to freely type in the name of the object they saw, instead of being given a forced-choice between two names (Kalmus, 1979, used a similar naming task with colors). Instead of administering the three orders within subjects,

[16]See Fisher (1967) for early examples of similar stimuli, much less controlled however than Stöttinger et al. (2016)'s. See Egré (2009) for philosophical perspectives on such examples concerning the relation between vagueness and perceptual ambiguity.

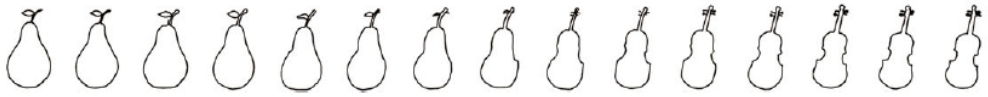

FIGURE 7. One of Stöttinger et al. (2016)'s 40 picture sets, morphing a pear into a violin.

the authors assigned one group of participants to the two ordered conditions, and assigned another group of participants to the random order condition. A comparison of the results across categories confirms that participants switch category earlier rather than later in the ordered conditions, although the underlying data is complex.[17] Unlike Egré et al. (2013), Stöttinger et al. did not run a separate perceptual matching task to see if the order effect was specific to linguistic categorization.

4.2.2. *Interpreting negative hysteresis.* Across the studies reviewed, we see that the phenomenon of negative hysteresis appears to be robust in cases of linguistic categorization. The phenomenon deserves emphasis, for part of the philosophical literature based on introspection had predicted positive instead of negative hysteresis in such ordered transitions (see Egré et al., 2013, for a review). How can the finding be explained?

Egré et al. (2013) review two possible explanations. One possible explanation would be to suppose, in agreement with the epistemic model of vagueness discussed in Bonini, Osherson, Viale, and Williamson (1999), that participants switch categories when they become uncertain of the membership of the stimulus, basically when they are within the margin of error needed to apply a category with confidence. The epistemic account presupposes that vague categories have a sharp but unknowable boundary, however, an assumption that is widely controversial. Egré et al. (2013) point out that the finding of negative hysteresis is compatible with a distinct model, in which categories are allowed to have multiple boundaries, but in which participants would basically switch category as soon as they step into the borderline area of competition between distinct verdicts.

A relevant model of the permissibility of multiple boundaries is proposed in Douven, Decock, Dietz, and Egré (2013). In this model, which bears some affinity with supervaluationism (see Douven, Wenmackers, Jraissati, & Decock, 2016), the vagueness of concepts is explained by the admission of multiple prototypes for a given category. Borderline cases

[17]In particular, although the overall data shows a tendency to switch early, more individual series show participants switching later than show them switching earlier, and most individual series do not show a significant difference in either direction. Moreover, the series showing the largest effect in either direction is one in which participants switch early, but there are some reasons to worry about this series. First, the size of the order effect this series shows is out of step with the other 39 series, coming in almost twice as large as the second-largest difference in either direction. Second, participants clearly struggled to decipher one of the endpoint images in this series, with 34% of participants unable to identify it (p. 205).

of a category correspond to cases that can be categorized in opposite ways, depending on which typical values have been selected to decide category membership. If we apply that model to the finding (something not done in Egré et al., 2013), the removal of an order effect in the perceptual condition could possibly be explained by the salience in perception, and thus in working memory, of *exactly two* specific prototypes for distinct categories. When such anchoring values are replaced by names of categories, participants behave as if they were sampling those anchoring values from a wider set, including more typical values than the two anchors of the perceptual task.

Note that this model by itself does not explain the direction of the effect (negative as opposed to positive hysteresis). However, Egré et al. (2013) point out that when participants are marched from a clear yellow to a clear orange, they start applying a category based on a strict membership criterion. Rationally, they should wish to signal a change as soon as they enter the borderline area, instead of continuing to apply the same name and change their criterion without notice. If the borderline area is spread around the middle stimulus in each series, then one would therefore expect them to switch earlier rather than later.

4.2.3. *Borderline cases: either, neither, or both?* Dynamic transitions involving morphed stimuli are also interesting to get a sense of whether stimuli in the middle range are likely to be perceived as belonging to *either* or even *both* of the end categories or whether they are more likely to be perceived as belonging to *neither*.

An interesting finding in Stöttinger et al. (2016) in this regard is that participants almost always correctly and unambiguously identified the first and last picture of each series under the expected name. For the middle objects in each series, there is an increase in the use of third names, but the use of a third name is generally marginal over the use of the names of the end stimuli (it peaks around 10% on average across the 40 sets). Closer inspection of their data reveals some variance across picture sets, however. In the series morphing a butterfly into a bowtie, for example, middle objects are hardly ever classified under a third category. In the series mapping a broom to a gun, the middle objects are more massively seen as being a third kind of object (such as a paddle, or a stick). Overall, however, what appears is that depending on the order of the transition, participants have no difficulty assigning the same stimulus to either of the opposing categories rather than to neither, as also demonstrated in Raffman's reversal cases.

In Egré et al. (2013), following work done by Ripley (2011) and Alxatib and Pelletier (2011), it was investigated whether and to what extent participants would agree to conjunctive descriptions of the form "both yellow and orange", and "both yellow and not yellow" to categorize middling stimuli in the color sets. Overall, what Egré et al. (2013) found was that ascriptions of the form "both P and Q" and "both P and not P" steadily increased and reached their maximum for the middle stimuli in each series, where they were used significantly more than either of their conjuncts. The finding confirms that participants have no difficulty recognizing that a borderline case can be assigned to either of the categories in those cases, but moreover that participants do not view the categories as exclusive of each other, but as admitting some overlap instead. Stoettinger et al.'s study could make us

think that it is an artifact: for it is not the same thing to categorize an object under *either* of two categories P and Q alternatively depending on the context, and to categorize it as being under a conjunctive category of the form "both P and Q". However, Stoettinger et al's study does not concern the ascription of adjectival categories, but of noun categories. A fair amount of general knowledge may block the inference from the observation that an object has two distinct features to the conclusion that it has both features at the same time. For example, a real-world object may not easily be categorized as "both a pair of scissors and a bodybuilder", even as its silhouette is likely to be perceived as sharing features of either kind of object, because we tend to envisage animate and inanimate objects as two exclusive categories in the first place, and no human bodybuilder could be thought of as a pair of scissors. Malt (1990) has established that animate categories in particular are mutually exclusive, while inanimate categories admit overlap. She asked participants to judge whether objects described as "halfway between" two animate or two inanimate categories (1) were probably one or the other category, (2) could be called either one, or (3) could not be part of either category, and found that participants predominantly chose the first response option for animate objects, but tended to choose the second response option for inanimate objects. For a more elaborate discussion regarding the use of conjunctions of (color) adjectives and their negation ("borderline contradictions", in the terminology of Ripley, 2011) we refer to Ripley (2011), Alxatib and Pelletier (2011), Serchuk, Hargreaves, and Zach (2011), and Egré and Zehr (2016). For broader considerations on the relation between vagueness and multistability in perception see Egré (2009) and Pelofi, de Gardelle, Egré, and Pressnitzer (2016).

5. Conclusions

We may draw two main conclusions from our review of the incidence of the sorites paradox in psychology. The first and maybe the most obvious one is that the sorites paradox can be operationalized and studied empirically. We have seen various ways in which sorites sequences have been created and studied, whether in a static or in a more dynamic setting. A second conclusion is that whenever this has been done, results are almost always accounted for in a probabilistic manner. Regarding that aspect, we can highlight four more specific points.

The first point, based on the discussion we gave of slippery slope arguments, is that soritical reasoning may be best viewed as a chain of propositions, the overall strength of which is dependent on the inductive strength of each conditional premise. The consideration of the probability of these individual conditional premises can serve to determine where in the argument a line should be drawn. A second and more specific lesson we can draw from the studies we reviewed is that the relation between discrimination and categorization is not as simple as the main premise of the sorites would suggest. Generally speaking, pairwise indiscriminability between two items does not guarantee that the items in question will always be placed under the same category, even as it makes that co-assignment highly probable.

Our third point is that small differences between items are always likely to make a difference, whether for discrimination or for categorization, even if those differences may not easily be detectable. In terms of discrimination, a small difference can nevertheless be such as to fall under the threshold picked to define what counts as a just noticeable difference, or be such as to go unnoticed altogether (as experiments on change blindness show us). In terms of categorization, a small difference generally increases or decreases the probability of assigning an item to a category, simply because the item ends up being a little more or a little less distant from the category prototype or anchoring values.

Finally, the tolerance principle is a risky principle, and the best way to represent the dependence of category membership on small differences is in terms of a principle of probabilistic closeness. This does not mean that the literal, non-probabilistic interpretation of the tolerance principle is worthless or not operative in how we judge and categorize. As the previous sections showed us, we do see assimilation and adaptation effects as a result of similarity. More generally, we see that context effects are numerous and subtle, and that the way in which items from a sorites series are presented has important consequences on either category maintenance or category switching.

Acknowledgments: The authors thank Elia Zardini and Sergi Oms for detailed and helpful comments. PE and SV thank the ANR Program TriLogMean ANR-14-CE30-0010-01 for funding, as well as grants ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL* for research carried out at the Department of Cognitive Studies of ENS. SV also thanks the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 313610.

References

Adams, E. W. (1998). A primer of probability logic.

Alxatib, S., & Pelletier, F. (2011). The psychology of vagueness: Borderline cases and contradictions. *Mind & Language*, *26*(3), 287–326.

Blackburn, S. (2002). *Being good: A short introduction to ethics*. OUP Oxford.

Bonini, N., Osherson, D., Viale, R., & Williamson, T. (1999). On the psychology of vague predicates. *Mind & language*, *14*(4), 377–393.

Borel, E. (1907). Un paradoxe économique: le sophisme du tas de blé et les vérités statistiques. *La Revue du Mois*, *4*, 688-699. (English translation by P. Égré and E. Gray, in *Erkenntnis* (79), 1081-1088, 2014 [An economic paradox: the sophism of the heap of wheat and statistical truths].)

Borel, E. (1950). *Probability and certainty*. Walker and Company: NY. (English translation 1963)

Bossuet, J. B. (1677). *Logique du dauphin*. Editions Universitaires. (1990 Edition by F. Laupiès)

Calder, A. J., Young, A. W., Perrett, D. I., Etcoff, N. L., & Rowland, D. (1996). Categorical perception of morphed facial expressions. *Visual Cognition*, *3*(2), 81–118.

Cayrol, N. (2016). Qu'est-ce qu'un sorite ? In *Mélanges en l'honneur du Professeur Jean Rossetto* (p. 241-264). LGDJ.

Cobreros, P., Egré, P., Ripley, D., & van Rooij, R. (2012). Tolerant, classical, strict. *The Journal of Philosophical Logic*, 1–39.

Corner, A., Hahn, U., & Oaksford, M. (2011). The psychological mechanism of the slippery slope argument. *Journal of Memory and Language*, *64*(2), 133–152.

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

Douven, I., Wenmackers, S., Jraissati, Y., & Decock, L. (2016). Measuring graded membership: The case of color. *Cognitive Science*. (doi: 10.1111/cogs.12359)

Dzhafarov, E. N., & Dzhafarov, D. D. (2012). The sorites paradox: A behavioral approach. *Qualitative mathematics for the social sciences: Mathematical models for research on cultural dynamics*, 105–136.

Egré, P. (2009). Soritical series and Fisher series. In A. Hieke & H. Leitgeb (Eds.), *Reduction: Between the mind and the brain* (pp. 91–115). Ontos Verlag.

Egré, P. (2011a). Perceptual ambiguity and the sorites. In R. Nouwen, R. van Rooij, U. Sauerland, & H. Schmitz (Eds.), *Vagueness in communication* (pp. 64–90). Springer.

Egré, P. (2011b). Review of Nicholas JJ Smith, *Vagueness and Degrees of Truth*. *Australasian Journal of Philosophy*, *89*(1), 177–180.

Egré, P. (2016). Vague judgment: a probabilistic account. *Synthese*. (doi:10.1007/s11229-016-1092-2)

Egré, P., & Barberousse, A. (2014). Borel on the Heap. *Erkenntnis*, *79*, 1043–1079.

Egré, P., De Gardelle, V., & Ripley, D. (2013). Vagueness and order effects in color categorization. *Journal of Logic, Language and Information*, *22*(4), 391–420.

Egré, P., & Zehr, J. (2016). Are gaps preferred to gluts? A closer look at borderline contradictions. In E. Castroviejo, G. Sassoon, & L. McNally (Eds.), *The semantics of gradability, vagueness, and scale structure - experimental perspectives.* Springer.

Fara, D. (2000). Shifting sands: an interest-relative theory of vagueness. *Philosophical Topics*, *28(1)*, 45-81. (Originally published under the name "Delia Graff")

Fechner, G. T. (1860). *Elemente der Psychophysik*. Breittsopf and Hartel.

Fisher, G. (1967). Measuring ambiguity. *The American Journal of Psychology*, *80 (4)*, 541–557.

Hahn, U., & Oaksford, M. (2006). A Bayesian approach to informal argument fallacies. *Synthese*, *152*(2), 207–236.

Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition*, *65*(2), 137–165.

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

Hampton, J. A., Estes, Z., Botbol, J., & Jaunbocus, Y. (2012). *Contrast and assimilation in simultaneous categorization.* (Unpublished Manuscript)

Hampton, J. A., Estes, Z., & Simmons, C. L. (2005). Comparison and contrast in perceptual categorization. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31 (6)*, 1459–1476.

Hardin, C. L. (1988). Phenomenal colors and sorites. *Noûs*, *22*, 213-234.

Harnad, S. (1987). Psychophysical and cognitive aspects of categorical perception: A critical overview. In *Categorical perception: The groundwork of cognition* (pp. 1–25). Cambridge University Press.

Horgan, T. (1994). Robust vagueness and the forced-march sorites paradox. *Philosophical Perspectives*, *8*, 159–188.

Kalmus, H. (1979). Dependence of colour naming and monochromator setting on the direction of preceding changes in wavelength. *British Journal of Physiological Optics*, *32(2)*, 1-9.

Lassiter, D., & Goodman, N. D. (2015). Adjectival vagueness in a bayesian model of interpretation. *Synthese*. (doi:10.1007/s11229-015-0786-1)

Lode, E. (1999). Slippery slope arguments and legal reasoning. *California Law Review*, 1469–1543.

Luce, R. D. (1956). Semiorders and a theory of utility discrimination. *Econometrica, Journal of the Econometric Society*, 178–191.

Luce, R. D. (1959). *Individual choice behavior*. Dover. (Reedition Dover 2005,)

Macmillan, N. A. (1987). Beyond the categorical/continuous distinction: A psychophysical approach to processing modes. In *Categorical perception: The groundwork of cognition* (pp. 53–85). Cambridge University Press.

Malt, B. C. (1990). Features and beliefs in the mental representation of categories. *Journal of Memory and Language*, *29*(3), 289–315.

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

O'Regan, J. K. (2001). Thoughts on change blindness. In *Vision and attention* (pp. 281–301). Springer.

O'Regan, J. K., Rensink, R. A., & Clark, J. J. (1999). Change-blindness as a result of 'mudsplashes'. *Nature*, *398*(6722), 34–34.

Peirce, C. S., & Baldwin, J. (1902). Sorites. In J. Baldwin (Ed.), *Dictionary of philosophy and psychology.*

Pelofi, C., de Gardelle, V., Egré, P., & Pressnitzer, D. (2016). Interindividual variability in auditory scene analysis revealed by confidence judgments. *Philosophical Transactions of the Royal Society B*. (http://dx.doi.org/10.1098/rstb.2016.0107)

Quine, W. V. (1970). Natural kinds. In N. Rescher (Ed.), *Essays in honor of carl g. hempel* (pp. 5–23). Dordrecht: Reidel.

Raffman, D. (1994). Vagueness without paradox. *Philosophical Review*, *103 (1)*, 41–74.

Raffman, D. (2011). Vagueness and observationality. In G. Ronzitti (Ed.), *Vagueness: A guide* (pp. 107–121). Springer.

Raffman, D. (2014). *Unruly words: A study of vague language.* Oxford University Press.

Repp, B. H. (1984). Categorical perception: Issues, methods, findings. *Speech and language: Advances in basic research and practice*, *10*, 243–335.

Ripley, D. (2011). Contradictions at the border. In R. Nouwen, H.-C. Schmitz, & R. van Rooij (Eds.), *Vagueness in communication* (pp. 169–188). Springer.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, *7*(4), 573–605.

Serchuk, P., Hargreaves, I., & Zach, R. (2011). Vagueness, logic and use: Four experimental studies on vagueness. *Mind & Language*, *26*(5), 540–573.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in cognitive sciences*, *1*(7), 261–267.

Smith, N. J. J. (2008). *Vagueness and degrees of truth.* Oxford: Oxford University Press.

Stöttinger, E., Sepahvand, N. M., Danckert, J., & Anderson, B. (2016). Assessing perceptual change with an ambiguous figures task: Normative data for 40 standard picture sets. *Behavior research methods*, *48*(1), 201–222.

van Rooij, R. (2011). Vagueness in linguistics. In G. Ronzitti (Ed.), *Vagueness: a guide* (p. 123-170). Springer.

Verheyen, S., De Deyne, S., Dry, M. J., & Storms, G. (2011). Uncovering contrast categories in categorization with a probabilistic threshold model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1515.

Verheyen, S., Dewil, S., & Egré, P. (2017). *Subjectivity in gradable adjectives: The case of tall and heavy.* (Manuscript submitted for publication)

Verheyen, S., & Egré, P. (2017). *Typicality and graded membership in dimensional adjectives.* (Manuscript submitted for publication)

Verheyen, S., Hampton, J. A., & Storms, G. (2010). A probabilistic threshold model: Analyzing semantic categorization data with the Rasch model. *Acta psychologica*, *135*(2), 216–225.

Verheyen, S., & Storms, G. (2011). Towards a categorization-based model of similarity. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd annual*

conference of the cognitive science society (pp. 614–619). Austin, TX: Cognitive Science Society.

Volokh, E. (2003). The mechanisms of the slippery slope. *Harvard Law Review*, *116*(4), 1026–1137.

Williams, B. (1985). Which slopes are slippery? In M. Lockwood (Ed.), *Moral dilemmas in modern medicine.*

Williamson, T. (1994). *Vagueness.* London: Routledge.

Wittgenstein, L. (1953). *Philosophische Untersuchungen.* John Wiley & Sons. (Translated into English by E. Anscombe, P.M.S. Hacker and J. Schulte)

Wright, C. (1976). Language mastery and the sorites paradox. In G. Evans & J. McDowell (Eds.), *Truth and meaning.* Oxford.