Diversity Effects in Subjective Probability Judgment

Constantinos Hadjichristidis

University of Trento

Janet Geipel

University of Chicago

Kishore Gopalakrishna Pillai

Amrita Vishwa Vidyapeetham

Author Note

Constantinos Hadjichristidis, Department of Economics and Management, University of Trento, Via Inama 5, Trento, 38122, Italy, E-mail: k.hadjichristidis@unitn.it, and Centre for Decision Research, Leeds University Business School, University of Leeds; Janet Geipel, Department of Psychology, University of Chicago, 5848 S. University Avenue, Chicago, IL, 60637, U.S.A., jgeipel@uchicago.edu; Kishore Gopalakrishna Pillai, Amrita School of Business, Amrita Vishwa Vidyapeetham, Coimbatore, India, kishorepillai@amrita.edu.

Correspondence concerning this article should be addressed to Constantinos Hadjichristidis, Department of Economics and Management, University of Trento, Via Inama 5, 38122 Trento, Italy. E-mail: k.hadjichristidis@unitn.it.

Word count: about 11200 (entire manuscript including title page, abstract, references, appendices, etc.)

Abstract

Previous research has shown that the judged probability of an event depends on whether its description mentions examples ("What is the probability that a randomly chosen Italian businessman will travel during the next month to Warsaw, Budapest, Prague or some other European city?") or does not mention examples ("What is the probability that a randomly chosen Italian businessman will travel during the next month to a European city?"). Here, we examined descriptions that mention examples and manipulated whether these are relatively similar (e.g., Warsaw, Budapest, Prague) or diverse (e.g., Warsaw, Marseilles, Helsinki). Four experiments (N = 1115) revealed a diversity effect: Overall, descriptions with diverse examples received higher probability judgments than descriptions with similar examples. We investigate possible mechanisms, and discuss theoretical and practical implications.

Keywords: probability judgment; support theory; similarity; diversity; coverage.

Diversity Effects in Subjective Probability Judgment

Decisions frequently depend on subjective assessments of probability. The decision of a person to get married, of a juror to cast a guilty or not guilty verdict, or of an international organization to impose or not to impose economic sanctions to a country, all depend to a certain extent on a subjective assessment of probability. Here we investigated how people assign probabilities to unpacked descriptions such as: "A randomly chosen Italian businessman will travel during the next month to Warsaw, Budapest, Prague or some other European city [as opposed to not travelling to any European city]"). Unpacked descriptions are descriptions that mention some or all examples of the target category. Specifically, we unpacked complex target categories, such as European cities, into similar examples and a residual category (e.g., Warsaw, Budapest, Prague or some other European city) or diverse examples and a residual category (e.g., Warsaw, Marseilles, Helsinki or some other European city). We predicted a diversity effect: All else being equal, descriptions with diverse examples would induce higher probability estimates than those with similar examples. Based on research we will discuss below, we expected that this would happen because descriptions with diverse examples prompt relatively fuller representations of the target categories.

Theoretical Background

Support Theory

According to the principle of *description invariance* (Tversky & Kahneman, 1981) or *extensionality* (Arrow, 1982), the way an event is described should not influence its judged probability. Empirical evidence, however, has revealed violations of this normative principle. For example, Rottenstreich and Tversky (1997) presented participants a brief scenario of a criminal trial. They asked one group (packed) to judge the probability "that the trial will not

result in a guilty verdict" while another group (unpacked) to judge the probability "of either a not guilty verdict or a hung jury rather than a guilty verdict." The median probability estimate of the unpacked group exceeded that of the packed group.

People's tendency to assign equal or greater probability to unpacked than to packed descriptions of a given category is known as *implicit subadditivity*. Tversky and colleagues considered it a central aspect of human judgment and developed *support theory* to explain it (Tversky & Koehler, 1994; Rottenstreich & Tversky, 1997). According to support theory, implicit subadditivity arises from memory retrieval and/or salience mechanisms. First, unpacked descriptions might remind judges of more members than co-extensional packed descriptions. In reference to the example above, the unpacked description might remind judges of the possibility of a hung jury, which may have slipped their minds when considering alternatives to a guilty verdict. The assumption is that each additional member a judge considers adds nonnegative evidence (support) for the target hypothesis, thereby increasing its judged probability. Second, even if the unpacked descriptions do not remind judges of additional category members, by virtue of being mentioned, the listed members gain salience. Their support increases and, through that, also the probability of the target hypothesis.

Narrow Interpretation Conjecture

Subsequent research by Sloman, Rottenstreich, Wisniewski, Hadjichristidis, and Fox (2004) challenged the ubiquity of implicit subadditivity (see also Hadjichristidis, Sloman, & Wisniewski, 2001; Hadjichristidis, Stibel, Sloman, Over, & Stevenson, 1999). This research examined unpacked descriptions that list a number of examples and include the rest in a catch-all, residual category. It found that certain unpacked descriptions, such as death from "pneumonia, diabetes, cirrhosis, or any other disease," lead to probability judgments that are

systematically *lower* than that of co-extensional packed descriptions, such as death from "a disease." To explain these and previous findings, Sloman et al. (2004) proposed the *narrow interpretation conjecture*. According to this conjecture, judges interpret packed and unpacked descriptions narrowly. They interpret packed descriptions in terms of their *typical* examples, that is, examples that are good representatives of the target category (Rosch & Mervis, 1975), and unpacked descriptions predominantly in terms of the unpacked examples. In contrast to support theory reminding explanation, the suggestion is that judges do not necessarily consider the unpacked examples *in addition* to those that the packed hypothesis naturally brings to mind, but perhaps *in place* of them.

Sloman et al. (2004) predicted that the effect of unpacking on probability judgment depends on two features of the unpacked examples: their typicality and support. If the unpacked examples are typical of the target category, that is, good representatives of it, then probability judgment would remain unaffected. These examples would match those that judges would spontaneously consider when presented with the packed description (cf. the salience explanation of support theory). But if the unpacked examples are atypical, that is, unrepresentative members of the target category, then they might influence probability judgment by replacing the more typical examples that individuals would have otherwise considered. Specifically, if the atypical examples offer comparatively low support in relation to the examples that they replace, then unpacking would *decrease* probability judgment. If they offer comparatively high support, then unpacking would *increase* probability judgment. Sloman et al. (2004) supported all these predictions empirically. These predictions have been also supported with other measures such as task completion estimates (Hadjichristidis, Summers, & Thomas, 2014) and spending estimates (Hadjichristidis, Pillai, & Burman, 2016), suggesting a domain general effect.

Present Research

Here, we aimed to extend the research on unpacking effects by investigating probability judgments assigned to different unpacked descriptions of a target category. Following Sloman et al. (2004), we used complex target categories and our unpacked descriptions included few examples and a catch-all residual hypothesis. The purpose was to make it difficult for the judges to think about all category members. Our objective was to examine whether, besides typicality and support (Sloman et al., 2004), the similarity or dissimilarity between the unpacked examples also affects judged probability. Consider the following unpacked descriptions of a target category, which we used in the present research:[1]

(1) A randomly chosen Italian businessman will travel during the next month to Warsaw, Budapest, Prague, or some other European city (as opposed to not travelling to any European city).

(2) A randomly chosen Italian businessman will travel during the next month to Warsaw, Marseilles, Helsinki, or some other European city (as opposed to not travelling to any European city).

The unpacked examples of the top and bottom descriptions were chosen such that they matched (in pairs) in terms of typicality and support (Warsaw matched Warsaw, Budapest matched Marseilles, and Prague matched Helsinki; for details of the pilot studies used to construct the descriptions, see Appendix A). The reason was to avoid the possibility that eventual findings would be due to differences in typicality or support. The difference between the descriptions is that whereas the unpacked examples of the top description are similar to one another because

[1] The original materials were in Italian. Here and throughout we present English translations.

they belong to the same salient subcategory (e.g., Eastern European cities), those of the bottom description are less similar as they belong to distinct subcategories (e.g., Eastern, Western, and Northern European cities).[2] We predicted a diversity effect: descriptions with diverse examples would prompt higher probability judgments than ones with similar examples.

Theoretical Motivation

Our prediction was motivated by the work of Osherson, Smith, Wilkie, Lopez, and Shafir (1990; see also Sloman, 1993) on category-based induction. In this paradigm, participants are presented with categorical arguments: a set of premises, which they must assume to be true, followed by a conclusion. Their task is to judge the strength of an argument, that is, the extent to which belief in its premises supports belief in its conclusion. Consider arguments A and B. The statements above the dotted line are the premises, and the one below the conclusion.

A.	B.
Tigers have a left aortic arch	Tigers have a left aortic arch
Lions have a left aortic arch	Chimps have a left aortic arch
Jaguars have a left aortic arch	Mice have a left aortic arch
---	---
Animals have a left aortic arch	Animals have a left aortic arch

Argument B is intuitively stronger than Argument A. Osherson et al. (1990) argued that this is because its premise categories—Tigers, Chimps and Mice—cover the conclusion category—

[2] Evaluations of similarity and category membership are dependent upon the knowledge base of the participants. We do not claim that description (1) contains objectively more similar members than description (2). Rather, that our participants perceived them as such. These descriptions might work differently for non-Europeans who might have different perceptions of similarity and category membership for these cities.

Animals—better than those of Argument A, by virtue of being more diverse. They defined semantic coverage as the similarity between the premise categories (e.g., Tigers, Chimps, Mice) and those that come to mind when considering the conclusion category (say, Cats, Monkeys, and Squirrels). The similarity process takes an exemplar of the conclusion category (e.g., Cats), compares it with each premise category (e.g., Cats-Tigers; Cats-Chimps; Cats-Mice), and registers the maximum similarity (e.g., the similarity between Cats and Tigers). The process continues with another exemplar of the conclusion category (e.g., Monkeys), and again it registers the maximum similarity (e.g., Monkeys-Chimps); and so on. Coverage is the average of these maximum similarities. It follows that the more diverse the premise categories are, the better they cover the conclusion category. Given a random exemplar of the conclusion category (e.g., Gorilla), it is more likely that it will have a close neighbour among the diverse (Tigers, Chimps, Mice) than among the similar premise categories (Tigers, Lions, Jaguars).

The Memory Retrieval Hypothesis. We predicted that semantic coverage might influence probability judgment by affecting how people represent the target category. Unpacked descriptions with diverse category elements might activate a fuller representation of the target category than descriptions with similar elements. We call this *the memory retrieval hypothesis*. Research suggests that category elements are likely to activate their close neighbours (e.g., Collins & Loftus, 1975). For example, Warsaw is likely to activate Budapest and perhaps Prague. Because of that, unpacking Warsaw, Budapest, and Prague, with respect to unpacking just Warsaw, might do little in terms of what category members get retrieved, or in which part of the target category a person focuses when forming a probability judgment. If anything, these elements might limit attention to the salient subcategory that includes them (e.g., Eastern European cities). However, when the elements belong to distinct subcategories, then they may

activate more category members and from more parts of the target category. Warsaw may activate Budapest, Marseilles Genoa, and Helsinki Stockholm. That is, unpacked descriptions with dissimilar elements might activate a broader representation of the target category.

Further supporting evidence comes from memory research using categorized lists (e.g., Banana, Apple, Orange, Mango; Chair, Table, Lamp, Couch; etc.). This research suggests that memory retrieval of listed members of a salient category (e.g., FRUIT) is all-or-none (Tulving & Psotka, 1971). If a person retrieves one item from it (e.g., Banana from FRUIT) then she is likely to retrieve most of the other items (e.g., Apple, Orange, and Mango). In relation to the present research, by unpacking members of distinct subcategories one would prompt the activation of other members of these subcategories. This supports that descriptions with diverse examples will lead to a fuller representation of the target category and thus trigger higher probability judgments than descriptions with similar examples.

The Misinterpretation Hypothesis. Diversity effects could also arise from pragmatic reasons (e.g., Grice, 1975). Unpacked descriptions with similar examples might promote lower probability judgments than ones with dissimilar examples because they lead judges to interpret the target category more narrowly (see Sloman et al., 2004; Van Boven & Epley, 2003). We call this *the misinterpretation hypothesis*. Consider the description with similar examples of the introductory example. Because the description only mentions Eastern European cities, a participant might interpret the question as asking for the probability that the Italian businessman will visit an *Eastern European city* (rather than *any European city*). In this case, finding lower judgments with these descriptions would not only be unsurprising but normative. In an effort to reduce the possibility of different category interpretations, we followed Sloman et al. (2004) and mentioned the alternative hypotheses (e.g., “What is the probability that a randomly chosen

Italian businessman will travel during the next month to Warsaw, Budapest, Prague or some other European city [*as opposed to not travelling to any European city*]?"). In Experiment 2 we examined directly how participants interpreted the target categories.

Overview of Experiments

Experiments 1a and 1b investigated the existence of a diversity effect. We presented participants with 12 unpacked descriptions, each referring to a different target category, and asked them to judge their probability. Participants assigned to the *diverse condition* received unpacked descriptions in which three relatively dissimilar category examples were unpacked, whereas those assigned to the *similar condition* received co-extensional descriptions in which three relatively more similar examples were unpacked. We predicted a higher overall probability judgment in the diverse condition.

Experiment 2 tested the misinterpretation hypothesis. We presented participants with either the similar or diverse versions of the 12 target descriptions, and asked them to assess whether some specific elements (similar to those of the diverse descriptions) belonged or did not belong to the target category. The misinterpretation hypothesis predicts that participants in the similar condition would not recognize some of these distant elements as members of the target category, as they would have interpreted the target category more narrowly.

Experiment 3 assessed the memory retrieval hypothesis via an item generation task and an ease-of-generation task. The item generation task involved asking participants to generate as many items as possible from the target category. We predicted that participants assigned to the diverse condition would generate items from more subcategories than participants assigned to the similar condition. The ease-of-generation task involved asking them to evaluate the ease of generating additional category members. With this task, we examined a variant of the memory

retrieval hypothesis according to which diversity influences the subjective ease of recalling further category items (see Schwarz et al., 1991; Schwarz & Vaughn, 2002), rather than the number or type of items recalled (for relevant evidence, see Biswas, Keller, & Burman, 2012).

Finally, Experiment 4 tested the diversity effect in the context of a lottery where there is low likelihood of misinterpretation. The aim was to test the robustness of the diversity effect.

Novelty

The present research is the first to examine whether diversity influences probability judgment in the context of unpacking effects. Although previous research examined how another category dimension— typicality—influences judged probability, typicality involves the relationship between *a single element* of a category (the unpacked example) and that category. The present research investigates the role of semantic coverage, which involves a relationship between *a group of category elements* (the unpacked examples) and that category (see Osherson et al., 1990). The present study is also one of few to test whether unpacking effects are driven by memory retrieval mechanisms (see also Tomlinson, 2007). In addition, it is the first to include a direct test for the misinterpretation hypothesis in the context of unpacking effects.

Experiment 1a

Experiment 1a (a pilot study) aimed to investigate the existence of a diversity effect. We asked one group of participants to judge the probability of versions with diverse examples of 12 events while another group that of versions with similar examples of the same 12 events. We predicted higher mean probability judgments for the diverse than the similar group. We also included a packed group, whose participants received versions of the 12 events in which no examples were mentioned. The purpose of the packed group was to provide a point of reference. However, we had no clear predictions of how the mean of this group would compare to those of

the other groups, as differences between packed and unpacked descriptions depend on the typicality and support of the unpacked examples (Sloman et al., 2004). Below we focus on our main prediction, and thus on the contrast between the similar and diverse groups.

Methods

The data of all studies are available on the Open Science Framework (OSF: https://osf.io/5smq8/?view_only=776ca41fecdd4a2191c442e531f488ed). Bayesian analyses are available by following the same link under Supplementary Results.

Participants

For this pilot study, we recruited participants online through university email lists. The study link remained active for 20 days. We recruited 86 Italian participants (73.2% females, 19.6% males, and 7.1% unspecified, M_{age} = 21.97 years, SD_{age} = 3.68, age range: 19 to 37 years). Thirty participants were randomly assigned to the unpacked *similar* condition, 26 to the unpacked *diverse* condition, and 30 to the *packed* condition.

Materials and Procedure

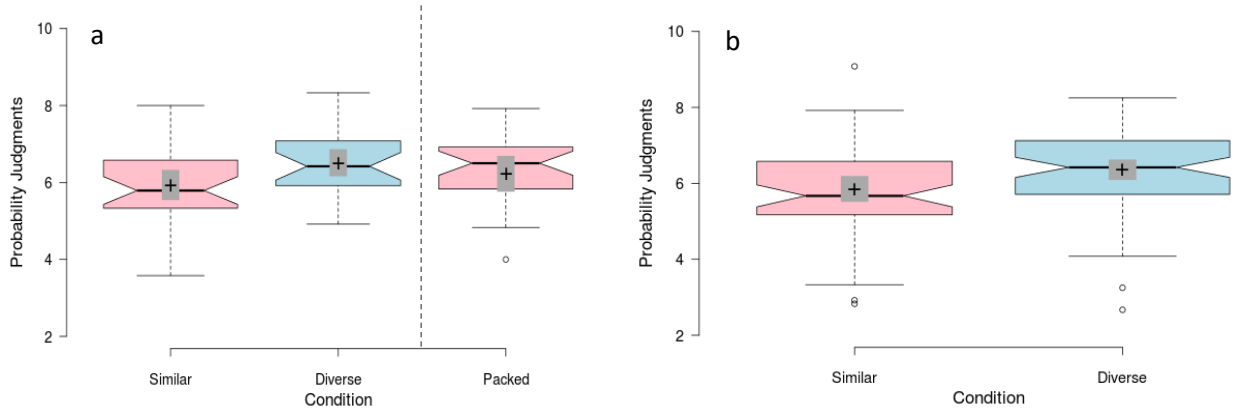
Participants were asked to judge the probability of 12 hypothetical events on an 11-point scale ranging from 0% = *impossible* (coded as 1) to 100% = *certain* (coded as 11). For example, participants in the packed condition were asked: "What is the probability that a randomly chosen person will go on vacation in a European country (as opposed not going on vacation in a European country)?" In the unpacked similar condition of this item the term 'European country' was replaced by 'Switzerland, Germany, Austria, or some other European country'. In the unpacked diverse condition, the corresponding term was 'Portugal, Ireland, Austria, or some other European country' (for the full text of the items, see Appendix A, Table A.1).

Results and Discussion

Figure 1a illustrates the mean probability ratings by condition. As predicted, the mean rating of the diverse condition was higher than that of the similar condition. The mean of the packed condition fell between those of the diverse and similar conditions.

Figure 1

Mean Probability Judgments by Diversity Condition (Experiment 1a and 1b)



Note. Plotted mean probability judgments across all 12 items by diversity condition in (a) Experiment 1a and (b) Experiment 1b. Centrelines illustrate the medians; box limits indicate the 25th and 75th percentiles, whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, and crosses represent sample means, grey bars indicate 95% confidence intervals of the means.

To test for a diversity effect, we submitted the probability ratings of the diverse and similar conditions to two simple one-way analyses of variance: one treating subjects as the random factor (F_1), and another treating item as the random factor (F_2). In the analysis by subjects, the mean of the diverse condition ($M_{\text{Diverse}} = 6.51$, $SD = 0.84$, 95% CI [6.17, 6.85]) was significantly higher than that of the similar condition ($M_{\text{Similar}} = 5.93$, $SD = 1.03$, 95% CI [5.55,

6.32]), Welch's $F_1(1, 53.87) = 5.34$, $p = .025$, $d_{Cohen} = 0.61$, 95% CI [0.07, 1.15]. A similar pattern was observed in the analysis by items (for an item-by-item presentation of means, see Appendix B, Table B.1). Unpacked descriptions with diverse items received higher probability ratings ($M_{Diverse} = 6.53$, $SD = 1.31$, 95% CI [5.70, 7.36]) than unpacked descriptions with similar items ($M_{Similar} = 5.94$, $SD = 1.45$, 95% CI [5.01, 6.87]), $F_2(1, 11) = 9.85$, $p = .009$, $d_{Cohen} = 0.91$, 95% CI [0.22, 1.58]. The predicted pattern of means was observed in 9 out of 12 items. Observing 9/12 or more successes by chance is less than .08 (two-tailed).

Experiment 1a served as a pilot study and provided initial support for the diversity effect: Descriptions listing diverse examples overall received a higher probability judgment than ones with similar examples.

Experiment 1b

Experiment 1b aimed to find further evidence for the diversity effect. The methods were similar to those of Experiment 1a with the exception that we dropped the packed condition.

Methods

Power Analysis

In order to determine the sample size needed for this study, we conducted an a-priori power analysis using the following estimates for a simple one-way ANOVA: $f = 0.31$ (based on Experiment 1a), $\alpha = .05$, power = .80, number of groups = 2. This analysis revealed a minimum sample of 84 participants.

Participants

We recruited participants online via university e-mail distribution lists. We recruited 131 participants (54.2% females, 45.8% males, $M_{age} = 30.02$ years, $SD_{age} = 14.10$, age range: 18 to 75

years). Seventy-one were randomly assigned to the diverse condition, and 60 to the similar condition.

Results and Discussion

Figure 1b illustrates the mean probability ratings by condition. The analysis by subjects yielded a significant main effect of condition, Welch's $F_1(1, 115.24) = 5.77$, $p = .018$, $d_{Cohen} = 0.43$, 95% CI [0.08, 0.78]. Participants in the diverse condition gave significantly higher probability ratings ($M_{Diverse} = 6.36$, $SD = 1.09$, 95% CI [6.08, 6.63]) than participants in the similar condition ($M_{Similar} = 5.86$, $SD = 1.27$, 95% CI [5.55, 6.16]). A similar pattern was observed in the analysis by items (for an item-by-item presentation of means, see Appendix B, Table B.1). Diverse descriptions received higher probability ratings ($M_{Diverse} = 6.36$, $SD = 1.47$, 95% CI [5.42, 7.29]) than similar descriptions ($M_{Similar} = 5.85$, $SD = 1.18$, 95% CI [5.11, 6.60]), $F_2(1, 11) = 11.60$, $p = .006$, $d_{Cohen} = 0.98$, 95% CI [0.27, 1.67]. This pattern of means was observed for 10 out of 12 items. Observing 10/12 or more successes by chance is less than .05 (two-tailed).

In Experiment 1b, we demonstrated the diversity effect on probability judgments with a larger sample of participants. Descriptions with diverse examples induced higher probability judgments than ones with similar examples.

Experiment 2

Experiment 2 tested whether participants interpreted the target categories of the diverse and similar descriptions differently. A fundamental difference between the memory retrieval and the misinterpretation hypotheses concerns a distinction between recall and recognition (see Tversky & Koehler, 1994). Extending on Tversky and Koehler (1994), we hold that the diversity effect is driven by recall. While evaluating the probability of a description with similar examples,

a participant may fail to access some category members that are dissimilar to those listed. Critically, however, given the opportunity, the participant would recognize such "distant" items as members of the target category. In contrast, the misinterpretation hypothesis predicts that participants would not recognize them even if asked explicitly, because they infer that the experimenter had a narrower category in mind.

Methods

Power Analysis

We used the same power analysis calculation as in Experiment 1b. The analysis revealed a minimum sample size of 84 participants.

Participants

We recruited 102 Italian participants (53.9% females, 46.1% males, M_{age} = 25.9 years, SD_{age} = 6.83, age range: 18 to 52 years) online through Prolific Academy (prolific.co). Fifty-five participants were randomly assigned to the diverse condition and 47 to the similar condition.

Materials and Procedure

Participants were instructed to imagine that a person, Paolo, made several predictions about another person, Maria. In total, we presented them with 12 different predictions, each linked to an item of Experiments 1a and 1b. Participants received all items in either its similar or diverse version. Here is an example from a similar version: *Paolo predicted that: "Maria went on vacation to Austria, Germany, Switzerland or some other European country (as opposed to not going on vacation in any European country)"*. The diverse version of this item mentioned "Austria, Ireland, Portugal". Following each prediction, on a separate page, participants saw five statements. Each contained information that either supported Paolo's prediction (e.g., "Maria went on vacation to Denmark") or did not support his prediction (e.g., "Maria went on vacation

to the United States"). For each statement, they were asked: "Does this information support Paolo's prediction?" (*Yes/No*).

As our objective was to test whether participants in the similar condition recognize distant elements as belonging to the target category, we created the target statements around the examples mentioned in the diverse condition (e.g., Austria, Ireland, and Portugal). One statement contained an example similar to the first unpacked element (e.g., Denmark [Austria]), another an example similar to the second element (e.g., Scotland [Ireland]), and another an example similar to the third element (e.g., Spain [Portugal]). A fourth statement contained an example that did not belong to the target category (e.g., the US), while a fifth statement contained an example that for some items belonged to the target category while for others it did not (e.g., for the European countries item, we included Kenya, which does not belong to the target category). Our aim was to avoid creating specific expectations (i.e., that for all items, participants should respond "Yes" to 4 statements and "No" to 1). Appendix C presents the 60 test items.

Due to the way the test statements were selected (to match the examples in the diverse condition), we increased the odds of finding more 'errors' in the similar condition. The order of presentation of the 12 items, and the order of presentation of the five statements underneath each item, were randomized separately for each participant.

Results and Discussion

We coded a response to an item as "1" (if the participant responded correctly to all 5 statements) or "0" (if the participant made one or more classifications errors). Next, for each participant, we calculated the sum score across the 12 items, which we used in subsequent analyses. In support of the misinterpretation hypothesis, a Mann-Whitney test indicated higher scores for the diverse condition ($M_{\text{diverse}} = 10.57$, $SD = 1.33$) than for the similar condition

($M_{similar}$ = 9.62, SD = 1.79), U = 859.0, p = .003, d_{Cohen} = 0.61, 95% CI [0.21, 1.00]. Similar results were obtained using a Welch's F-test, $F(1, 98.22) = 9.53$, p = .003, d_{Cohen} = 0.61, 95% CI [0.21, 1.00].

Next, we conducted separate chi-square tests for each item. We observed significant misinterpretation for three items: present, major and family (see Appendix B). For all three, there were more classification errors in the similar than in the diverse condition: present (2.1% vs. 21.8%), χ^2 (1, N = 102) = 8.84, p = .003, φ = .29; major (19.1% vs. 38.2%), χ^2 (1, N = 102) = 4.42, p = .035, φ = .21; family (6.4% vs. 38.2%), χ^2 (1, N = 102) = 14.24, p < .001, φ = .37.

Re-Analyses of Experiments 1a and 1b

In light of these findings, we revisited the analyses of Experiments 1a and 1b. The aim was to examine whether the diversity effect is explained by misinterpretation. For each participant, we computed two mean probability judgments: one across the items for which we found significant misinterpretation and one across the items for which we did not. Condition (diverse, similar) and Item type (misinterpretation, no misinterpretation) did not interact with Experiment (1a, 1b), $F(1, 182) = 0.08$, p = .774, $\eta^2 < 0.01$. Hence, we combined the two datasets. We found a main effect of condition, $F(1, 184) = 12.15$, p = .001, $\eta^2 = 0.06$. Critically, we did not find a Condition × Item type interaction, $F(1, 184) = 2.06$, p = .153, $\eta^2 = 0.01$. We also found a main effect of item type, $F(1, 184) = 33.94$, p < .001, $\eta^2 = 0.16$.

Experiment 2 tested the misinterpretation hypothesis. Although we found that three items were interpreted more narrowly in the similar than the diverse condition, these interpretation differences did not explain the diversity effect of Experiments 1a and 1b. The diversity effect was present and comparable across items independent of whether or not they showed interpretational differences.

Experiment 3

Experiment 3 tested the memory retrieval hypothesis. Following Schwarz and colleagues (e.g., Schwarz et al., 1991; Schwarz & Vaughn, 2002), we considered two possibilities. First, we considered the possibility that semantic coverage acts by influencing the subjective ease of recall, that is, the subjective experience of ease or difficulty associated with recalling or imagining further category exemplars (see Biswas et al., 2012). Specifically, diverse as opposed to similar descriptions might make it easier for judges to generate further category exemplars. To examine this possibility, we included an ease-of-generation task. After the probability judgment task, we asked participants to think about other category members (without listing them) and to rate how easy or difficult they found this to be.

Second, we considered the possibility that coverage acts through the content of recall, that is, the type of exemplars that come to mind. Descriptions with diverse as opposed to similar examples might remind judges of exemplars from a greater number of salient subcategories, which would be evidence for a fuller representation of the target category. To examine this possibility, we included an item generation task. We asked participants to list as many items of the category as they could. For each participant, we computed the number of generated items (excluding those mentioned in the description), and, critically, also the number of subcategories the generated items spanned. We expected that participants in the diverse condition would generate items that span a greater number of different subcategories.

Methods

The study's design, hypothesis, sample size, and planned analyses was preregistered on AsPredicted.org (https://aspredicted.org/blind.php?x=xe7bf5).

Power Analysis

In order to determine the sample size for this study, we conducted an a-priori power analysis using the following estimates for a simple one-way ANOVA: $f = 0.20$ (minimum effect size from previous experiments), $\alpha = .05$, power = .99 (to be conservative), number of groups = 2. This analysis revealed a minimum sample of 462 participants. We requested 480 participants online through Prolific Academy (prolific.co) to account for possible exclusions.

Participants

We received responses from 496 Italian participants (47.4% female, 51.0% male, 1.6% other, $M_{age} = 23.4$ years, $SD_{age} = 3.62$), 16 more than we requested. Following the pre-registration, we excluded one participant for failing an attention check and another for not generating any items. The results below are based on the remaining 494 participants (47.4% female, 51.0% male, 1.6% other, $M_{age} = 23.4$ years, $SD_{age} = 3.63$, age range: 18 to 51 years). Of these, 249 participants were randomly assigned to the similar condition and 245 participants to the diverse condition.

Materials and Procedure

Participants were presented with one item for which we did not observe significant interpretation issues—either the Car or the Illness item (see Appendix B)—in either its similar or diverse form. We considered using the Vacation and Business trip items, but decided against it given the current travel restrictions due to COVID-19. Participants were first asked to estimate the probability of the target event on a 11-point scale ranging from 0% = *impossible* (coded as 1) to 100% = *certain* (coded as 11). Next, they were presented with the ease-of-generation task ("How easy was it for you to think about [different text depending on condition]?", 1 = *Not at all easy* to 7 = *Extremely easy*) and the item generation task, in which they were asked to generate as

many examples of the target category as possible within 45 seconds (the countdown was visible to participants). The presentation order of these tasks was randomized across participants. Finally, to check our diversity manipulation, we asked participants to judge the similarity or diversity of the listed examples ("How similar or diverse were [different text depending on condition]", 1 = *Very similar* to 7 = *Very different*).

Results

Manipulation Check

Validating our manipulation, the examples in the diverse condition were judged as more different ($M = 5.24$, $SD = 1.40$) than those in the similar condition ($M = 3.94$, $SD = 1.38$), Welch's $F(1, 491.57) = 107.18$, $p < .001$, $d_{Cohen} = 0.93$, 95% CI [0.75, 1.12].

Probability

Figure 2 illustrates the results. Participants in the diverse condition gave significantly higher probability estimates ($M = 7.59$, $SD = 2.20$; 95% CI [7.31, 7.87]) than participants in the similar condition ($M = 7.05$, $SD = 2.49$, 95% CI [6.74, 7.36]), Welch's $F(1, 486.54) = 6.53$, $p = .011$, $d_{Cohen} = 0.23$, 95% CI [0.05, 0.41].

Ease-of-generation

Participants in the diverse condition gave similar ratings ($M = 4.09$, $SD = 1.85$, 95% CI [3.85, 4.32]) as participants in the similar condition ($M = 3.94$, $SD = 2.04$, 95% CI [3.69, 4.20]), Welch's $F(1, 488.77) = 0.66$, $p = .417$, $d_{Cohen} = 0.07$, 95% CI [–0.10, 0.25].

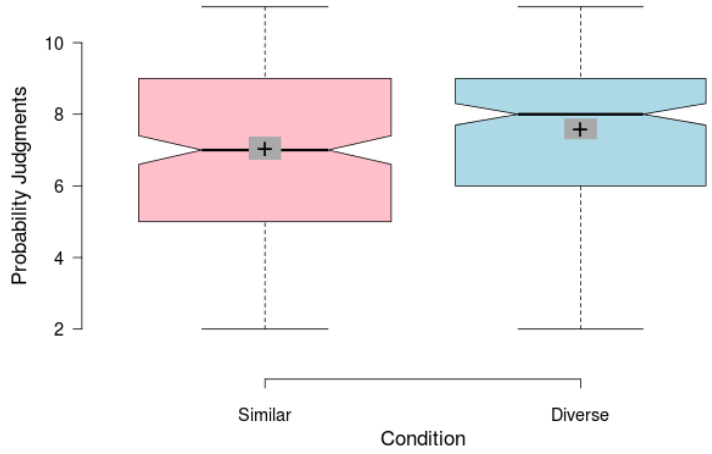
Item Generation and Number of Categories

We first analysed the number of generated items excluding those mentioned in the description. Then, we classified the generated items into categories and counted the number of categories the items spanned. The categories used to classify the items were those that emerged

from the pilot study, which were also used to generate the examples for the diverse and similar descriptions. A research assistant who was unaware of the purpose of the experiment performed the classification of items into categories. During coding, several issues emerged. For the illness item, together with illnesses participants often generated symptoms (e.g., fever). Similarly, for the car item, besides specific models participants often generated car manufacturers (Alfa Romeo). For number of items, the mention of an illness symptom or car manufacturer was treated as a separate item. For number of categories, such mentions were ignored because most symptoms are associated with multiple illnesses and most manufacturers offer models in several different categories.

Figure 2

Mean Probability Judgments by Diversity Condition (Experiment 3)



Note. Centrelines illustrate the medians, box limits indicate the 25th and 75th percentiles, whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, and crosses represent sample means, grey bars indicate 95% confidence intervals of the means.

Participants in the diverse condition generated a comparable number of items (M = 5.44, SD = 2.41, Md = 5.00) as participants in the similar condition (M = 5.24, SD = 2.19, Md = 5.00), Welch's $F(1, 485.48) = 0.93$, $p = .335$, $d_{Cohen} = 0.09$, 95% CI [–0.09, 0.26], ($H[1] < 0.59$, $p = .442$; $\chi^2[1, N = 494] = 1.14$, $p = .287$). However, participants in the diverse condition generated items from more categories (M = 2.99, SD = 1.14, Md = 3.00) than participants in the similar condition (M = 2.76, SD = 1.17, Md = 3.00), Welch's $F(1, 484.53) = 4.84$, $p = .028$, $d_{Cohen} = 0.20$, 95% CI [0.02, 0.38].

Correlations between Probability, Ease-of-generation, Number of Items and Number of Categories

We conducted Spearman's rho correlations. There was a significant positive correlation between probability judgment and ease-of-generation, $r(492) = .153$, $p = .001$; the higher the ease-of-generation ratings, the higher the probability judgments. The was no significant correlation between probability judgment and number of generated items, $r(492) = .085$, $p = .060$. There was a significant negative correlation between probability judgment and number of categories, $r(485) = –.146$, $p = .001$; the fewer categories participants accessed, the higher the probability judgment. This result was surprising as it goes against the memory retrieval hypothesis. We followed it up by conducting separate correlations for the two items. Neither correlation was significant (Illness: $r[245] = .04$; Car: $r[238] = .02$). The source of the negative correlation in the overall analysis was because the illness item prompted a higher number of categories, while the car item induced higher probability judgments.

Focusing on the predictors, there was a significant positive correlation between ease-of-generation and number of items, $r(492) = .304$, $p < .001$; the higher the ease-of-generation the greater the number of items generated. There was a significant positive correlation between ease-

of-generation and number of categories, $r(485) = .232$, $p < .001$; the higher the ease-of-generation, the greater the number of categories the participants accessed. Finally, there was a significant positive correlation between generated items and number of categories, $r(485) = .523$, $p < .001$; the more items participants generated the more categories they accessed.

Does Ease-of-generation, Number of Items Generated and Number of Categories Mediate the Diversity Effect on Probability Judgments?

We examined whether ease-of-generation, number of items and number of categories mediate the effect of diversity on probability judgments. As outcome variable, we used probability judgment, and as predictor diversity condition (0 = Low, 1 = High). As preregistered, we conducted a mediation analysis testing for an indirect effect using 10,000 bootstrapped samples and the 95% confidence interval (Hayes, 2017). We found no significant indirect effect (total indirect effect: $b = –0.07$, 95% CI [–0.206, 0.061]; direct effect: $b = 0.66$, 95% CI [0.255, 1.063]). Hence, these variables did not mediate the effect of diversity on probability judgments.

Discussion

Experiment 3 provides further evidence for the diversity effect, but no indication of the underpinning mechanism. We found no differences between the diverse and similar conditions in terms of how easy participants found it to think of category examples (ease-of-generation) or how many examples participants generated (item generation). The only significant difference was that participants in the diverse condition generated items from more categories. However, mediation analyses showed that none of these three variables explains the diversity effect.

Experiment 4

Experiment 2 showed that participants interpreted the target category of some descriptions with similar examples more narrowly than that of descriptions with diverse

examples. The three problematic cases all involved ambiguous categories (What counts as a "scientific field", "present for a mother", or an "organized family activity?"). The problem, therefore, might be category ambiguity rather than misinterpretation. It is possible that the diverse examples encouraged a broader definition of the target category, the similar examples encouraged a narrower interpretation, while both definitions fall within the legitimate bounds that ambiguity in definition allows for. Category ambiguity might also apply to the other categories mentioned in the descriptions such as a "randomly chosen person". However, it is not clear how the unpacked examples could systematically influence the interpretation of such other categories. In Experiment 4, we aimed to replicate the diversity effect with a simple lottery task that has a clear structure and therefore eventual category ambiguity is unlikely to influence judgments.

Methods

Power Analysis

In order to determine the sample size needed for this study, we conducted an a-priori power analysis using the following estimates for a simple one-way ANOVA: $f = 0.20$ (minimum observed effect size), $\alpha = .05$, power = .80, number of groups = 2. This analysis revealed a minimum sample of 200 participants.

Participants

We recruited 297 Italian participants (44.1% female, 54.2% male, 1.7% other, M_{age} = 26.5, SD_{Age} = 6.77, age range: 18 to 60 years) online through Prolific Academy (prolific.co). Of them, 150 were randomly assigned to the diverse condition and 147 to the similar condition.

Materials and Procedure

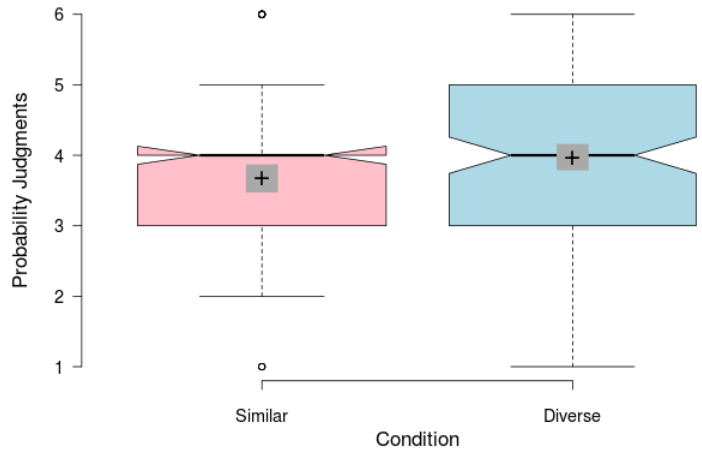
Participants were asked to imagine a lottery with tickets numbered from 1 to 100. They further had to imagine a person having bought all lottery tickets with numbers greater than 50. Participants in the similar condition were asked to estimate the probability that this person will win the lottery with ticket number #70, #72, #74 or any other of their lottery tickets. Participants in the diverse condition were asked to estimate the probability that this person will win the lottery with ticket number #57, #74, #89 or any other of their lottery tickets. Participants were asked to respond on a 6-point scale ranging from 1 (*Not at all probable*) to 6 (*Very probable*).

Results and Discussion

Figure 3 illustrates the results. The mean probability judgment was higher in the diverse condition ($M = 3.98$, $SD = 1.11$) than in the similar condition ($M = 3.68$, $SD = 1.19$), Welch's $F(1, 294.39) = 5.03$, $p = .026$, $d_{Cohen} = 0.26$, 95% CI [0.03, 0.49].

Figure 3

Mean Probability Judgments by Diversity Condition (Experiment 4)



Note. Centrelines illustrate the medians, box limits indicate the 25th and 75th percentiles, whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, and crosses represent sample means, grey bars indicate 95% confidence intervals of the means.

To assess the robustness of this finding, we also conducted a Mann-Whitney *U*-test, $U = 12{,}585.5$, $p = .027$, $r = 0.14$, 95% CI [0.01, 0.27]. In sum, in Experiment 4 we demonstrated the diversity effect using a simpler lottery task where category ambiguity is less likely.

General Discussion

Previous research suggests that the judged probability of an unpacked hypothesis depends on the typicality and support of the unpacked examples (Sloman et al., 2004). Here, we show that it also depends on the degree to which the unpacked examples provide semantic coverage for the target category. In Experiments 1a and 1b we examined diverse and similar unpacked versions of 12 complex events whose examples matched in terms of typicality and probability. Overall, the diverse versions were assigned a higher probability than the similar versions. Experiment 2 found that three of the 12 target categories were interpreted more narrowly in the similar than in the diverse condition. However, a re-analysis of Experiments 1a and 1b showed that such interpretation differences did not drive the diversity effect. Furthermore, in Experiment 4 we demonstrated the diversity effect in a simple lottery context that has a clear structure, and therefore eventual category ambiguities are unlikely to influence judgment.

In Experiment 3, we examined the diversity effect with two items that did not show interpretation issues, and investigated its underpinning mechanism. We found a diversity effect, but could not pinpoint the mechanism. Participants in the diverse and similar conditions generated a comparable number of items, and found it equally easy to think about other category items. The only significant difference was that participants in the diverse condition generated items from more categories than did participants in the similar condition. However, a mediation analysis showed that none of these factors—number of generated items, ease-of-generation, number of categories—explains the association between diversity and probability judgment.

Theoretical Implications

The present findings cannot be explained by support theory (Tversky & Koehler, 1994; Rottenstreich & Tversky, 1997) or the narrow interpretation conjecture (Sloman et al., 2004), because neither of them considers semantic coverage. The narrow interpretation conjecture comes close with typicality, but typicality and coverage are distinct. Typicality refers to the similarity between *a single member* of a category and the target category, whereas coverage to the similarity between *a group of members* of a category and the category.

From a broader perspective, coverage, just like typicality, originates in the literature on category representation and use. Following the work of Sloman et al. (2004), the present research builds a novel bridge between categorization and judgment and decision-making research. There have been several efforts to import concepts of categorization to the domain of judgment and decision making. Mental accounting (e.g., Thaler, 1999), category-bound thinking (Sunstein, Kahneman, Schkade, & Ritov, 2002; see also Bonini, Graffeo, Hadjichristidis, & Ritov, 2019), construal level theory (Trope & Liberman, 2010), the representativeness heuristic (e.g., Kahneman & Tversky, 1972), the prototype heuristic (Kahneman & Frederick, 2002), and the editing phase of prospect theory (Kahneman & Tversky, 1979), are all prominent examples of such attempts. The present findings show that further links are possible.

What Drives the Diversity Effect?

Sticking with the memory retrieval hypothesis, one possibility is that coverage acts in a different manner to what we envisaged. Following support theory and the narrow interpretation conjecture, we have assumed that individuals represent target categories via specific exemplars. That is, we work within an exemplar-based view of category representation (e.g., Medin & Schaffer, 1978). However, alternative views of representing categories exist such as the feature-

based view (e.g., Sloman et al., 1993) and the frame-based view (e.g., Minsky, 1975). For example, in the frame-based view categories are represented by slots (e.g., GEOGRAPHIC LOCATION) and filler values (e.g., Eastern Europe). It could be that descriptions with diverse examples prompt category representations whose slots contain more filler values (e.g., Eastern Europe, Western Europe, Northern Europe) than ones with similar examples (e.g., Eastern Europe). This could explain the diversity effect, but also why number of items or categories did no mediate the association between condition and probability: because the generated items are not a direct representation of how judges instantiated the target categories.

Moving away from the memory retrieval hypothesis, it could be that diversity acts more directly. For example, diversity might affect probability judgments via (semantic) proximity. In the process of evaluating the probability of a hypothesis, a judge might bring to mind examples that would make it true. Whatever examples a judge considers, there is a higher chance that they will have a close neighbour among those in the diverse group than among those in the similar group (see Osherson et al., 1990). In sum, it could be that judges evaluate probability via a proximity heuristic (see Teigen, 1998, 2005), which returns higher values for descriptions with diverse examples.

A third possibility is that the similarity or diversity of a group of members affects how typical (representative) the group is of the target category. Coverage can be interpreted as a generalized measure of typicality between a group of category members and a salient superordinate category. In this view, typicality is just a limiting case where the group consists of one member (see Osherson et al., 1990). Following this line of thought, the diversity effect could be due to the representativeness heuristic. A group of diverse examples is more representative of

the target category than is a group of similar examples and this difference in representativeness might drive the higher probability judgments in the diverse condition.

Limitations and Future Directions

One limitation of the present research concerns the operationalization of coverage. We asked participants to divide categories into subcategories and, in the case of multiple salient classifications, we constructed similar and diverse sets of examples that were robust to the multiple classifications (see Appendix A). Future studies could employ different procedures to measure coverage such as the one used by Osherson et al. (1990). Their procedure entails asking participants to generate as many exemplars as possible from a target category, and then to judge the similarity between all possible pairs of exemplars. From the pairwise similarity estimates, one can compute the coverage that different sets of examples offer for the target category. One could also input these pairwise similarity ratings into a number of statistical models that output a representation of the semantic space of the target category (for help with deciding which technique to use, see Pothos & Chater, 2001). Such representations of a category could help develop further hypotheses. Could it be that the density of the category space close to the unpacked elements also influences probability judgment? Other possibilities for operationalizing coverage include semantic relatedness measures based on widely available software such as WordNet::Similarity (Pedersen, Patwardhan, & Michelizzi, 2004).

A second limitation of the present research, and of previous research on unpacking effects, is that it does not offer direct evidence for an explanation. Future research could investigate the hypothesis that diverse descriptions prompt a fuller representation of the target category through tasks like the lexical decision task (e.g., Meyer & Schvaneveldt, 1971). One could provide participants either with the diverse or similar description and then ask them to

decide as fast and as accurately as possible whether some letter strings do or do not represent words (Yes/No). Participants should be faster to respond that distant examples belong to the category given the diverse description than the similar one. Future research could also devise tasks to test whether the diversity effect is due to a proximity or representativeness heuristic.

Practical Implications

Often we have limited space or time to describe complex categories. Consider a scientific paper, an advertisement, or a presidential candidate speech. What elements should we unpack to promote our idea or product? The present findings suggest that we should aim for diversity (see also Kim & Keil, 2003). A scientific paper proposing a novel effect might have more chances of being persuasive if it demonstrates it with three different studies (from different domains) than with three similar studies (from the same domain). To promote an all-inclusive travel insurance policy, one could unpack diverse situations in which one might need the insurance (illness, bad weather, terrorist threat, strike) rather than similar ones (different types of illnesses; see also Johnson, Hershey, Meszaros & Kunreuther, 1993). Similarly, to increase the effectiveness of a fitness campaign one could unpack diverse benefits (e.g., health, psychological, and social), while to increase the effectiveness of an anti-smoking campaign one might unpack diverse drawbacks (e.g., health, psychological, and financial). In general, the present findings suggest that diversity could be used strategically to nudge citizens towards beneficial activities or away from harming ones. These recommendations are novel; they cannot be derived from support theory or the narrow interpretation conjecture.

Conclusion

The judged probability of a hypothesis that lists some examples and includes the rest in a residual category is influenced by the extent to which the examples mentioned are diverse; the

extent to which they cover the target category. Controlling for typicality and probability, the greater the number of distinct subcategories the listed examples spanned, the higher the judged probability. This finding offers a novel link between categorization and judgment under uncertainty. It suggests that judged probability not only depends on the attributes of the listed examples, such as how typical each example is of the target category, but also on properties of the examples as a set. The present findings could inform communication strategies to promote beneficial behaviors or discourage harming ones.

Acknowledgements

We thank Alessia Dorigoni, Ambra Ferrari and Jacopo Slanzi for research assistance.

References

Arrow, K. J. (1982). Risk perception in psychology and economics. *Economic Inquiry, 20*, 1-9.

Biswas, D., Keller, L. R., & Burman, B. (2012). Making probability judgments of future product failures: The role of mental unpacking. *Journal of Consumer Psychology 22*, 237–248. doi:10.1016/j.jcps.2011.03.002

Bonini, N., Graffeo, M., Hadjichristidis, C., & Ritov, I. (2019). Category-bounded emotional enhancement: Spillover effects in the valuation of public goods. *Cognition and Emotion, 33* (7), doi: 10.1080/02699931.2018.1559802

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335–359. doi: 10.1016/S0022-5371(73)80014-3

Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic priming. *Psychological Review, 82*, 407-428. doi: 10.1037//0033-295X.82.6.407

Dougherty, M. R. P., & Hunter, J. E. (2003a). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica, 31*, 263–282. doi: 10.1016/S0001-6918(03)00033-7

Dougherty, M. R. P., & Hunter, J. E. (2003b). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition, 31*, 968–982. doi:10.3758/BF03196449

Grice, H. P. (1975). Logic and Conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics* (pp. 41–58). New York, US: Academic Press.

Hadjichristidis, C., Pillai, K. G., & Burman, B. (2016, May). *Effects of unpacking in spending predictions: The role of typicality.* Paper presented at the 44th AMS (Academy of marketing Science) Annual Conference, Lake Buena Vista, Florida.

Hadjichristidis, C., Sloman, S. A., & Wisniewski, E. (2001). Judging the probability of representative and unrepresentative unpackings. In Johanna D. Moore (Ed.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 376–380). Mahwah, NJ: Erlbaum.

Hadjichristidis, C., Stibel, J., Sloman, S. A., Over, D. E., & Stevenson, R. J. (1999). Opening Pandora's box: Selective unpacking and superadditivity. In Sebastiano Bagnara (Ed.), *Proceedings of the European Conference on Cognitive Science* (pp. 185–190). Siena, Italy.

Hadjichristidis, C., Summers, B., & Thomas, K. (2014). Unpacking estimates of task duration: The role of typicality and temporality. *Journal of Experimental Social Psychology, 51*, 45–50. doi: 10.1016/j.jesp.2013.10.009

Johnson, E., Hershey, J., Meszaros, J. & Kunreuther, H. (1993). Framing, probability distortions and insurance decisions. *Journal of Risk and Uncertainty, 7*, 35–51. doi: 10.1007/BF01065313

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman, (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). New York, US: Cambridge University Press.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3 (3),* 430–454. doi: 10.1016/0010-0285(72)90016-3.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An analysis of decision under risk. *Econometrica, 47*, 263–291. doi: 10.2307/1914185.

Kim, N. S., & Keil, F. C. (2003). From symptoms to causes: Diversity effects in diagnostic reasoning. *Memory & Cognition, 31*, 155–165. doi: 10.3758/BF03196090

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification leaming. *Psychological Review, 85*, 207–238. doi: 10.1037/0033-295X.85.3.207

Meyer, D.E., & Schvaneveldt, R.W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*(2), 227–234. doi: 10.1037/h0031564

Minsky, M. (1975). A framework for representing knowledge. In P.H.Winston (Ed.), *The psychology of computer vision*. New York: McGraw-Hill.

Murphy, G. L. (2003). *The big book of concepts.* Cambridge, MA: MIT Press.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*, 185–200. doi: 10.3758/s13415-013-0221-3

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). Wordnet:: similarity — measuring the relatedness of concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence* (AAAI-04). AAAI Press, Cambridge, MA, pages 1024–1025.

Pothos, E. M., & Chater, N. (2001). Categorization by simplicity: A minimum description length approach to unsupervised clustering. In U. Hahn & M. Ramscar (Eds.), *Similarity and categorization* (pp. 51–72). Oxford, England: Oxford University Press.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573–605. doi: 10.1016/0010-0285(75)90024-9

Rottenstreich, Y., and Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review, 104*, 203–231. doi: 10.1037/0033-295X.104.2.406

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology*, 61, 195–202. doi: 10.1037/0022-3514.61.2.195

Schwarz, N., & Vaughn, L. A. (2002). The availability heuristic revisited: Ease of recall and content of recall as distinct sources of information. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases* (pp. 103–119). New York: Cambridge University Press.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology, 25*, 231–280. doi:10.1006/cogp.1993.1006

Sloman, S. A., Rottenstreich, Y., Wisniewski, C., Hadjichristidis, C., and Fox, C. R. (2004). Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 573–582. doi: 10.1037/0278-7393.30.3.573

Sunstein, C.R., Kahneman, D., Schkade, D., & Ritov, I. (2002). Predictably incoherent judgments. *Stanford Law Review, 54*, 1153-1215. doi: 10.2139/ssrn.279181

Teigen, K. H. (1998). When the unreal is more likely than the real: Post hoc probability judgements and counterfactual closeness, *Thinking & Reasoning*, *4*(2), 147–177. doi: 10.1080/135467898394193

Teigen, K. H. (2005). The proximity heuristic in judgments of accident probabilities. *British Journal of Psychology, 96*, 423–40. doi: 10.1348/000712605X47431.

Thaler, R. H. (1999). Mental accounting matters. *Journal of Behavioral Decision Making, 12 (3*), 183-206. doi: 10.1002/(SICI)1099-0771(199909)12:3<183::AID-BDM318>3.0.CO;2-F

Tomlinson, T. D. (2007). *The role of part set cuing and retrieval induced forgetting in subjective probability judgments* (Master's thesis). Retrieved from http://drum.lib.umd.edu/handle/1903/7266.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117(2)*, 440-463. doi: 10.1037/a0018963

Tulving, E., & Psotka, J. (1971). Retroactive inhibition in free recall: Inaccessibility of information available in the memory store. *Journal of Experimental Psychology*, *87*(1), 1–8. doi: 10.1037/h0030185

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 21*, 453-58. doi: 10.1126/science.7455683

Tversky, A., & Koehler, D. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*, 547–567. doi: 10.1037/0033-295X.101.4.547

Van Boven, L., & Epley, N. (2003). The unpacking effect in evaluative judgments: When the whole is less than the sum of its parts. *Journal of Experimental Social Psychology, 39*, 263–269. doi: 10.1016/S0022-1031(02)00516-4

Appendix A

Methods and Results of Pretests

Pretests

Typicality. Nineteen University of Trento students voluntarily participated in an online-based study. Participants were presented with the 12 target categories each followed by a number of items. Each category was presented on a separate page. Participants had to rate the typicality of each member to a target category on a 7-point goodness-of-example scale ranging from 1 (not at all a good example) to 7 (the best example). Goodness-of-example is a standard measure of typicality (e.g., Murphy, 2003; Rosch & Mervis, 1975). For example, participants were presented with the category "European cities as potential business trip destinations of a randomly chosen Italian businessman." Underneath it, in a vertical manner, they were shown a list of city names including London, Milan, Paris, and Budapest. Next to each city name (on its right) participants were presented with the 7-point scale, and had to respond by ticking the appropriate number. To be able to proceed to the next page, participants had to respond to all items in a given page.

Probability. A new sample of 21 University of Trento students voluntarily participated in an online-based study. Participants were asked to judge the probability of the 12 target hypotheses each time substituting the target category with a different member. They had to assign a rating on a 7-point scale ranging from 1 (extremely unlikely) to 7 (extremely likely). The presentation of the categories and items were similar to the typicality pretest. For example, participants were instructed: "Please imagine a randomly chosen Italian business man who has a business trip in a European city in the next month. How likely is it that the Italian business man chooses to go to…" Underneath this description, in a vertical manner, they were presented with various cities such as London, Milan, Paris, and Budapest. Next to each city they were presented

with the 7-point probability scale. Their task was to respond by ticking the appropriate number in that scale.

Coverage. The students that participated in the typicality and probability pre-tests (N = 40), as a final task, were presented once more with the categories followed by the associated examples. For each category, the participants were instructed to first read carefully all of the listed examples and then divide the category into three or more distinct subcategories. They were provided with 5 empty boxes in which they could write the names of the subcategories. As an example, they were given a list of animals including Lion, Elephant, Boa, and Trout, and as possible subcategories: Mammals, Birds, Reptiles, and Fish. Unsurprisingly, in this pre-test participants subdivisions varied. For example, for the item concerning dogs (see Table A.1, item 10) a frequent sub-division was in terms of size (Small, Medium, Large) while another in terms of type (Guard dogs, Hunting dogs, Family dogs). When this happened, we selected similar and diverse examples in such a way that they were robust across the different classifications. For example, in relation to the dog item, the diverse condition included three dogs of different sizes and types, while the similar condition it included three small dogs that are typically viewed as family dogs.

On the basis of the results from the typicality, probability, and coverage pre-tests, for each target category, we selected three elements for the similar description (A1, A2, A3) and three for the diverse description (A1, B1, C1) such that: (1) The elements of the similar description were from the same sub-category, whereas those of the dissimilar description were from different subcategories (see Table A.1); (2) The elements A2—B1 and A3—C1 were closely matched in terms of typicality and probability (see Table A.2). Note that for A1 this criterion is automatically satisfied as both conditions shared this item.

Table A.1

Descriptions of the Items Used in Experiments 1a and 1b.

Item	Description	Similar examples A1, A2, A3	Dissimilar examples A1, B1, C1
1	What is the probability that a randomly chosen person will go on vacation to _ _ _ or some other European country [as opposed to not going on vacation in a European country].	Austria, Germany, Switzerland	Austria, Ireland, Portugal
2	What is the probability that a randomly chosen young person is practicing _ _ _ or some other sport [as opposed to that the young person not practicing any sport].	karate, judo, kick boxing	karate, ice skating, horse riding
3	What is the probability that a randomly chosen child would buy as a present for his or her mother on mother's day _ _ _ or some other present [as opposed to that the child not buying any present for her].	necklace, earrings, watch	necklace, CD, chocolate
4	What is the probability that a randomly chosen 25-years old has a degree in _ _ _ or in some other scientific field [as opposed to not having a degree in a scientific field].	biology, pharmacology, medicine	biology, psychology, mathematics
5	What is the probability that a randomly chosen person would buy _ _ _ or some other product at Feltrinelli [as opposed to not buying anything at Feltrinelli]. (*Feltrinelli is similar to Barnes & Nobles).	travel guide, historical atlas, world atlas	travel guide, e-reader, gift card
6	What is the probability that a randomly chosen Italian businessman will travel during the next month to _ _ _ or some other European city [as opposed to not travelling to any European city].	Warsaw, Budapest, Prague	Warsaw, Marseilles, Helsinki
7	What is the probability that a randomly chosen Italian family would either buy or receive as a present _ _ _ or some other item from Ikea [as opposed to neither buying nor receiving any present from Ikea].	Drinking glasses, cups, plates	Drinking glasses, lampshades, towels
8	What is the probability that a randomly chosen 40-year old Italian man owns _ _ _ or some other car model [as opposed to not owning any car].	Fiat Panda, Lancia Ypsilon, Peugeot 107	Fiat Panda, Toyota RAV4, VW Passat
9	What is the probability that a randomly chosen person suffered during the last six months from _ _ _ or some other illness [as opposed to not suffering from any illness].	asthma, bronchitis, pneumonia	asthma, dermatitis, neurosis
10	What is the probability that a randomly chosen Italian family owns _ _ _ or some other type of dog [as opposed to not owning a dog].	Chihuahua, Pug, Poodle	Chihuahua, Dalmatian, German shepherd
11	What is the probability that a randomly chosen family would organise _ _ _ or some other activity to spend a day together [as opposed to not organising any activity].	visit an art exhibition, a museum, a castle	visit an art exhibition, shopping, board games
12	What is the probability that a randomly chosen person would order in a restaurant _ _ _ or some other alcoholic beverage [as opposed to not ordering any alcoholic beverage].	dark beer, lager beer, double malt beer	dark beer, white wine, grappa

Table A.2

Mean Goodness-of-Example and Probability Judgments for the Items Used in Experiments 1a and 1b.

	Description	Goodness-of-example			Probability		
Item	A1, A2, B1, A3, C1	A1	A2, B1	A3, C1	A1	A2, B1	A3, C1
1	Austria, Germany, Ireland, Switzerland, Portugal	4.89	4.94, 5.00	4.78, 5.72	4.09	4.78, 4.22	3.52, 3.83
2	karate, judo, ice skating, kick boxing, horse riding	4.06	3.94, 3.56	3.83, 3.78	4.05	3.86, 3.62	3.62, 3.67
3	necklace, earrings, CD, watch, chocolate	5.17	4.94, 4.44	4.17, 5.17	5.08	4.58, 3.83	3.33, 4.21
4	biology, pharmacology, psychology, medicine, mathematics	6.22	5.33, 5.22	5.67, 5.89	5.04	4.54, 4.71	5.17, 4.96
5	travel guide, historical atlas, e-reader, world atlas, gift card	-- *	4.33, 4.33	4.39, 4.17	-- *	3.38, 3.62	3.67, 3.41
6	Warsaw, Budapest, Marseilles, Prague, Helsinki	4.11	3.61, 3.67	4.39, 4.72	2.76	3.08, 3.20	3.20, 3.36
7	Drinking glasses, cups, lampshades, plates, towels	5.61	4.72, 5.17	5.11, 4.78	4.81	4.24, 4.90	4.81, 4.24
8	Fiat Panda, Lancia Ypsilon, Toyota RAV4, Peugeot 107, VW Passat	4.72	5.00, 4.67	4.78, 4.89	4.76	4.43, 4.24	4.67, 4.67
9	asthma, bronchitis, dermatitis, pneumonia, neurosis	5.17	5.72, 5.44	4.67, 4.44	4.92	5.00, 5.08	4.44, 4.36
10	Chihuahua, Pug, Dalmatian, Poodle, German shepherd	4.67	4.56, 4.56	5.94, 5.89	4.63	4.29, 3.50	5.04, 5.25
11	visit an art exhibition, visit a museum, shopping, visit a castle, board games	4.78	4.94, 4.89	5.28, 5.33	3.61	4.39, 4.57	4.65, 4.57
12	dark beer, lager beer, white wine, double malt beer, grappa	5.00	6.22, 6.61	4.83, 4.44	4.71	5.83, 5.96	5.00, 4.67

*Note.** For item 5, we found it hard to construct diverse and similar examples versions. Because of this, we ended up using a shared example (travel guide) that was not pre-tested. Hence, we do not have goodness-of-example and probability ratings for this item.

Appendix B

Mean Probability Judgments by Items of Experiments 1a and 1b.

Table B.1

Mean Probability Judgments by Item (1—12) and Experiment (1a—1b).

		Probability judgments			
		Experiment 1a		Experiment 1b	
Item	Description	$M_{Diverse}$	$M_{Similar}$	$M_{Diverse}$	$M_{Similar}$
1	Vacation	**6.08**	**5.76**	**4.81**	**4.51**
2	Sport	**7.31**	**7.24**	**6.30**	**6.03**
3	Present	6.58	6.79	**5.83**	**5.17**
4	Major	**5.69**	**4.67**	**3.99**	**3.46**
5	Bookstore	**4.12**	**3.67**	2.44	2.71
6	Business trip	**6.12**	**5.37**	**4.75**	**4.51**
7	Ikea	**5.88**	**4.53**	4.51	5.00
8	Car	**9.56**	**8.60**	**8.18**	**7.03**
9	Illness	**7.62**	**6.60**	**6.13**	**5.17**
10	Dog	6.42	6.63	**5.10**	**4.10**
11	Family	**6.00**	**4.28**	**5.35**	**4.49**
12	Alcohol	6.96	7.13	**6.90**	**6.07**

Note. In boldface pairs of means that are consistent with our hypothesis (i.e., $M_{Diverse} > M_{Similar}$).

Appendix C

The Elements Used to Construct the Target Statements of Experiment 3.

Table C.1

The Elements Used to Construct the Target Statements of Experiment 3 by Item (1—12).

Item	Description	Elements
1	Vacation	Denmark, Scotland, Spain, **United States**, **Kenya**
2	Sport	Mixed martial arts, ski, polo, **guitar lessons**, **piano lessons**
3	Present	Ring, DVD, pralines, **book (for father)**, perfume
4	Major	Medicine, Cognitive Sciences, Physics, **Art**, Chemistry
5	Bookstore	Tourist guide, e-book, store coupon, **fruits**, **ice-cream**
6	Business trip	Moscow, Paris, Stockholm, **New York**, Berlin
7	Ikea	Tableware, table lamp, bed sheets, **CD from Media World**, carpet
8	Car	Mini, VW Touareg, BMW 730, **Vespa**, **Ducati**
9	Illness	Emphysema, herpes, depression, **no illness**, ear infection
10	Dogs	Maltese, French pointing dog, Belgian shepherd, **Siamese cat**, Neapolitan Mastiff
11	Family	Photo exhibition, mall, Monopoly, **cinema (alone)**, theater
12	Alcohol	Red beer, red wine, herbal liqueur, **Coca-Cola**, Campari

Note. In boldface elements that do not support Paolo's prediction. Note that some descriptions have one such element (e.g., items 3 and 6) while others have two (e.g., items 1 and 8).