Semantics & Pragmatics Volume 8, Article 13: 1–45, 2015
http://dx.doi.org/10.3765/sp.8.13

Moral asymmetries and the semantics of *many* [*]

Paul Egré
Institut Jean-Nicod

Florian Cova
University of Geneva

Submitted 2014-04-08 / Accepted 2014-06-01 / Revision received 2014-10-17 / Final version received 2015-01-13 / Published 2015-10-07

Abstract

We present the results of four experiments concerning the evaluations people make of sentences involving *many*, showing that two sentences of the form *many As are Bs* and *many As are Cs* need not be equivalent when evaluated relative to a background in which B and C have the same cardinality and proportion to A, but in which B and C are predicates with opposite semantic and affective values. The data provide evidence that subjects lower the standard relevant to ascribe *many* for the more negatively valued predicate, and that judgments involving *many* are sensitive to moral considerations in a broad sense, namely to expectations involving a representation of the desirability as opposed to the mere probability of an outcome. We relate the results to similar semantic asymmetries discussed in the psychological literature, in particular to the Knobe effect and to framing effects.

[*] This paper extends and supersedes a preliminary report on this research, which appeared under the same title in the *Proceedings of the Amsterdam Graduate Philosophy Conference 2011*. We are indebted to Louise McNally and three reviewers of *Semantics and Pragmatics* for very helpful comments, which substantially improved our work, as well as to reviewers on an earlier version. We also thank audiences in Amsterdam, Paris, Geneva, UCLA, London, Princeton and MIT. Special thanks go to Asaf Bachrach, Pierre Barrouillet, Heather Burnett, Emmanuel Chemla, Mikaël Cozic, Inès Crespo, Cressida Gaukroger, Nat Hansen, Markus Kneer, Edouard Machery, Robert May, David Ripley, Robert van Rooij, Galit Sassoon, Stephen Schiffer, Philippe Schlenker, Benjamin Spector, Ed Stabler, Brent Strickland and Frank Veltman for valuable exchanges and suggestions, and to Josh Knobe and Seth Yalcin for earlier exchanges on this topic. This work was done with the support of the ANR project "Cognitive origins of vagueness" (ANR-07-JCJC-0070), as well as the project PETAF (Perspectival Thoughts and Facts), and the CRNAP network (Cooperative Network in Analytic Philosophy). We also thank the Department of Cognitive Studies at ENS (IEC grants ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL*), the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement Ni324115–FRONTSEM (PI: Schlenker), and the NCCR Affective Sciences financed by the Swiss National Science Foundation (no. 51NF40-104897) and hosted by the University of Geneva.

Keywords: *many*, quantification, context-sensitivity, vagueness, Knobe effect, framing effects, severity bias, probability, value, evaluativity, moral norms, prospect theory

1 Introduction

The aim of this paper is to investigate the semantics of the quantifier *many* in relation to a family of moral asymmetries that have been documented in various places in the literature.

The first of those is an asymmetry evidenced by Knobe 2003 regarding people's ordinary judgments about intentional action. Knobe presented the following scenario to two groups of subjects. One group read the scenario with the word *harm*, the other group with *help* uniformly in place of *harm*:

> The vice-president of a company went to the chairman of the board and said, "We are thinking of starting a new program. It will help us increase profits, but it will also {harm / help} the environment." The chairman of the board answered, "I don't care at all about {harming / helping} the environment. I just want to make as much profit as I can. Let's start the new program." They started the new program. Sure enough, the environment was {harmed / helped}.

Subjects in each group then had to respond by yes or no to the following question:

(1) Did the chairman intentionally {harm / help} the environment?

In the *harm* condition, a large majority of subjects agreed that the chairman intentionally harmed the environment. In the *help* condition, by contrast, most subjects denied that the chairman intentionally helped the environment. This effect is surprising, since in each scenario, the chairman exerts the same influence on the outcome, and is described as equally informed and equally indifferent toward the side-effect.

In a recent paper, Pettit & Knobe 2009 have outlined a convincing explanation for this asymmetry, based on an analogy with the semantics of gradable adjectives. What Pettit and Knobe point out is that two liquids, coffee and beer, can be at the same temperature — say, 20°C — but be such that one would judge the first to be cold and the second not to be cold. To judge that the coffee is cold is to judge that it is colder than it should be, given the expected temperature for coffee; to deny that the beer is cold is to judge that the the beer is not as cold as it should be, given the expected temperature for beer. Phrased in terms of degrees, to judge that coffee is cold is to judge that the degree to which it is cold exceeds the norm or standard relevant to ascribe coldness to coffee; to deny that beer is cold is to judge that the degree to which it is cold is below the norm relevant to beer. By analogy, to judge that an action type is done intentionally is to judge that the degree of intentionalness attached to the action is above the normative threshold relevant for that kind of action.[1] In the same way in which the threshold for *cold* can vary from beer to coffee, the threshold for *intentional* can thus vary from *harm* to *help* along the dimensions relevant in Knobe's scenario. Thus, although the chairman's internal properties and causal influence are the same in each condition, whether an action is described as *harm* or *help* makes different standards of comparison salient in order to judge whether that action was done intentionally.

Further evidence was proposed in Egré 2014 to articulate and substantiate Pettit and Knobe's explanation. Basically, the suggestion is that the Knobe effect might be an instance of a more general asymmetry concerning our expectations between negatively valued versus positively valued outcomes (see the summary and discussion in Egré 2010). This asymmetry has been documented in the psychological literature concerning people's perception of risk in particular, starting with Tversky and Kahneman's experiments on framing effects (Kahneman & Tversky 1979, Tversky & Kahneman 1981), and including what Weber & Hilton 1990 describe as a "worry effect". In a recent study on risk communication, Pighin, Bonnefon & Savadori 2011 compared the rankings given by four groups of pregnant women, concerning the probability of {1 in 307 / 1 in 28} that a particular child will have {insomnia / Down Syndrome}, on a 7-point scale ranging from "extremely low" to "extremely high". What they found is that subjects ranked significantly higher the lower probability of 1 in 307 for the child to have Down Syndrome, in comparison to the probability of 1 in 28 for the child to have insomnia. Those answers were found to correlate with how severe they judged each disease to be. Similar results were obtained by Krumpal et al. 2011 in a set of studies concerning people's perception of victimization risks. What they observed is that with

[1] We talk of "intentionalness" rather than "intention" since data by Knobe show that speakers distinguish between *harming/helping intentionally* and *intending to harm/help*. See Egré 2014 for more on the semantics of *intentional* proper.

more severe offenses (bill-dodging versus robbery versus murder), “identical verbal likelihood expressions correspond to lower numerical probabilities” (Krumpal et al. 2011: 1343). This phenomenon, also known as the “severity bias” (see Pighin, Bonnefon & Savadori 2011, Bonnefon & Villejoubert 2006, Krumpal et al. 2011), suggests that the same mechanism operates in judgments about whether an action is intentional and in judgments about whether a probability is high. In the latter case, two identical probability values on the scale from 0 to 1 can be such that the first will be judged to be high in comparison to the standard relevant for a severe outcome, while the other will be judged not to be high in comparison to the standard relevant for a non-severe outcome. Contrasts in judgment correspond to shifts of the standard relevant in each domain.

In Egré 2014, the hypothesis was originally formulated that one should observe essentially the same kind of asymmetry in judgments about quantities expressed in terms of the vague quantifier *many*. That is, the prediction was made that judgments should be found to differ in pairs of the form *many As are Bs* versus *many As are Cs*, where *B* and *C* are contradictory predicates with different affective values attached to them, even as the number of *A*s that are *B*s is the same as the number of *A*s that are *C*s. In what follows, we present the results of four experiments intended to test this prediction. In each experiment, we first asked people to assent or dissent to such a pair of sentences for a specific scenario in order to probe their truth-conditional intuitions. In a second phase, we looked for information about people’s positioning of the threshold in relation to their judgments about *many*. One advantage of this methodology is that it fits the way in which judgments involving gradable expressions can be modeled, essentially in terms of a comparison to an implicit normative standard (see Sapir 1944, Bartsch & Vennemann 1972, Fara 2000, Lappin 2000, Kennedy 2007, Solt 2011). Furthermore, because the scale associated with *many* is more transparent than the one for *intentional*, the data give us insight into the way in which norms and expectations determine our judgments involving vague predicates.

The sensitivity of quantifier words like *many* to prior expectations is a well-known phenomenon, and has been the object of several studies (see in particular Moxey & Sanford 2000 for a review). For example, Moxey & Sanford 1993 investigated the understanding of vague quantifiers like *few* and *many* in relation to distinct base-rate expectations. Likewise, Fernando & Kamp 1996 propose an account of *many* based on probabilities: *many As are Bs* is true basically when the probability of there being fewer *AB*s than there are, given how many *A*s there are, is greater than a threshold. In this paper, our perspective is related, but differs in that we seek to isolate the impact of what we might call moral expectations (in a sense to be made precise), as opposed to mere probabilistic expectations, on the quantifier *many*. The second, third and fourth experiments we present, in particular, propose a way of contrasting those two notions of expectations.[2]

In the next section, we start with some background on the semantic analysis of the severity bias and the Knobe effect, in terms of shifting standards of comparison, and extend this analysis to sentences of the form *many As are Bs*. In Section 3, we present the results of our first experiment and show that the data comport with this semantic model. Section 4 presents a follow up experiment, in which probabilistic expectations were properly controlled for. Experiment 1 and Experiment 2 only concern the same pair of predicates, namely *survive* versus *die*. In Sections 5 and 6, we present the results of two further experiments intended to test less loaded pairs of predicates, namely *stay* versus *leave* (in relation to a party), and *pass* versus *fail* (in relation to an exam). Each time, the second element of the pair refers to an outcome that is undesirable in comparison to the first. Both pairs replicate the asymmetry obtained initially. In Section 7, we draw some more general lessons from those findings regarding the opposition between two kinds of expectations (probabilistic versus moral), and conclude with some considerations about our data's relationship with Kahneman and Tversky's prospect theory on the one hand, and about the link between Knobe-type asymmetries and framing effects on the other.

2 Shifting standards and the semantics of *many*

2.1 The severity bias and the Knobe effect

Pighin, Bonnefon & Savadori's (2011) data on judgments about probabilities in relation to the scale involving the predicates *high* and *low* indicate that (2a) can be judged true and (2b) false in the same context without inconsistency:

(2) a. A probability of $1/307$ for a child to have Down Syndrome is high.
    b. A probability of $1/28$ for a child to have insomnia is high.

[2] For the distinction between moral and non-moral expectations in relation to Knobe-type asymmetries, see in particular Mandelbaum & Ripley 2010 and Cova, Dupoux & Jacob 2012.

The contrast between the two judgments can be represented by means of a semantics *à la* Bartsch & Vennemann (1972), assuming that *high* is a predicate that maps individuals to degrees, and that for a probability to be high is for (the degree attached to) that probability to be higher than the norm for highness in relation to the kind of event under consideration.[3] Let p denote the probability $1/307$ of having Down Syndrome and p' denote the probability $1/28$ of having insomnia, and compare:

(3) a. $[\![\textit{high}]\!]_w(p) \geq \mathbf{norm}_w(\textit{DownSyndrome})(\textit{high})$
  b. $[\![\textit{high}]\!]_w(p') \geq \mathbf{norm}_w(\textit{insomnia})(\textit{high})$

Assume for simplicity that $[\![\textit{high}]\!]_w(p) = 1/307$ and $[\![\textit{high}]\!]_w(p') = 1/28$, or in other words that that the degrees of highness in the context w are identical to the numerical probabilities, but that $\mathbf{norm}_w(\textit{DownSyndrome})(\textit{high}) = 1/1000$ and that $\mathbf{norm}_w(\textit{insomnia})(\textit{high}) = 1/15$. This is a situation in which (3a) is true and (3b) false.

In agreement with Pettit & Knobe's (2009) remarks, essentially the same analysis can be given for Knobe's examples based on the adjective *intentional*. Thus, one can consistently judge that both (4a) and (4b) are true, assuming the standard of comparison for whether an action type is *intentional* is set lower for *harming the environment* than for *helping the environment* on the relevant scale of comparison.

(4) a. The chairman's harming of the environment was intentional.
  b. The chairman's helping of the environment was not intentional.

Let h stand for *the chairman's harming of the environment* and h' for *the chairman's helping of the environment*, and assume that $[\![\textit{intentional}]\!]_w(h) = [\![\textit{intentional}]\!]_w(h')$, or in other words that the degrees of intentionalness attached to each action type are identical. Letting the standard shift from one case to the other, it is possible to have:

(5) a. $[\![\textit{intentional}]\!]_w(h) \geq \mathbf{norm}_w(\textit{harm})(\textit{intentional})$
  b. $[\![\textit{intentional}]\!]_w(h') < \mathbf{norm}_w(\textit{help})(\textit{intentional})$

[3] We use large inequalities instead of strict inequalities in what follows (*high* means 'at least as high as', *many* means 'at least as many as'). Nothing essential hinges on it here and we could have used strict inequalities as far as we can see. We also use the symbol $>$ instead of Kennedy's $\succ$, which Kennedy intends for 'significantly more than'. Finally, we identify the degree attached to how high a probability is with that probability, though that assumption is a psychological simplification.

Prima facie, the Pighin pair and the Knobe pair suggest that the more detrimental an event type is perceived to be, the lower the standard will tend to be positioned for interest-relative predicates such as *high* or *intentional*. This hypothesis calls for some qualifications, however.

A first caveat concerns the fact that the validity of the hypothesis depends on the polarity of the adjective under consideration. Obviously, for the negative adjective *unintentional*, one would expect the threshold to be lower for *help* than for *harm* — that is, lower for the beneficial outcome in this case (and similarly, *mutatis mutandis*, if *low* were used instead of *high* to qualify probabilities). A second issue, of more methodological nature, concerns the fact that neither Knobe's experiment nor Pighin's experiment provides us with much information about how subjects locate the thresholds relative to each other in either condition. In the case of the adjective *intentional* as applied to action types, the problem is intrinsically more complex than for *high* as applied to probabilities, since the structure of the associated scale of comparison is not transparent in this case. But even for *high*, Pighin et al.'s study does not allow us to see how far the thresholds will be located apart from each other depending on the kind of disease under consideration.

To get information of that kind, the four experiments we designed presented participants with a pair of sentences involving the vague quantifier *many* as applied to a count noun, in order to have a simple and discrete scale of comparison: the scale of natural numbers with their usual ordering. Before getting to the details, we first rehearse a few basic facts about the semantics of *many* in sentences of the form *many As are Bs*.

2.2 *Many*

For the most part, the semantics of the quantifier expression *many* obeys the same pattern as that already introduced for the gradable adjective *high*. On the approach we adopt here, to say that *many As are Bs* is to consider that there are more *AB*s than what is expected, normal, or comparatively standard in a given context (see Sapir 1944, Keenan & Stavi 1986, Lappin 2000, Solt 2009, Solt 2011).[4] As in the case of gradable adjectives like *high*,

[4] By this disjunctive formulation, we try to be as general as possible regarding the truth conditions of *many As are Bs*. As pointed out to us, not all uses of *many As are Bs* can be paraphrased as *there are more ABs than expected*, if we consider examples such as *many American families go on vacation during the summer months*. But the latter might be paraphrased, depending on the context, as *more American families go on vacation during the summer months than* {*during other months / in other countries / ...*}. See in particular Solt

the threshold for *many* is vague and context-dependent. Moreover, *many* is not purely extensional, but is intensional in the sense that *many As are Bs* and *many Cs are Ds* may differ in truth-value even if A is coextensional with C, and B with D (see Keenan & Stavi 1986, Fernando & Kamp 1996, Lappin 2000). This means that the threshold for *many* can be sensitive to the meaning of its arguments, namely to which comparison class is specified by its restrictor (its first argument) as well as by its nuclear scope (its second argument). To take an example given by Lappin, (6a) can be judged false and (6b) true in a situation in which the violinists are all the musicians, and the women are all the Italians present:

(6) a. Many musicians at the concert are women.
    b. Many violinists at the concert are Italian.

Suppose that there are 100 musicians at the concert, all violinists, including 30 Italian women, and 70 non-Italian men such that no more than five of them have the same nationality. Then (6a) may be judged false if the threshold for women to count as many musicians when there are 100 musicians is 50, and (6b) may be judged true if the threshold for Italians to count as many violinists when there are 100 violinists is 20.[5]

By analogy with the truth-conditions given earlier for *high*, we propose that *many As are Bs* is true in context w provided:[6]

(7) $|A \cap B|_w \geq \mathbf{norm}_w(A, B, |A|_w)$

This says that many As are Bs provided the actual number of ABs is greater than the relevant norm or expected value for As and Bs relative to the actual cardinality specified by the restrictor A. As assumed for the semantics of gradable adjectives given above, this normative value varies depending on further contextual elements relevant in w besides the three main arguments

2009, Solt 2011 and Greer 2014 for an analysis of *many* wholly in terms of explicit or implicit comparison class arguments, intended to capture both extensional and intensional readings of *many*. The truth conditions we lay out for *many*, while inspired by Lappin 2000, may actually be restated in the terms of either Solt's or Greer's comparison class analyses. We think that an intensionalized version of Greer's account, in particular, could accommodate our results. We return to this comparison in Section 7.3, Footnote 21.

[5] See Moxey & Sanford (1993: 74) for a similar illustration of the context-sensitivity of judgments involving *many* for identical numbers and proportions. Their example involves *having a cold* vs. *enjoying the party* relative to *children in the class* vs. *people at a party*.

[6] We write $|A|_w$ instead of the more accurate $|[\![A]\!]_w|$ to denote the cardinality of the set of A relative to w, and similarly $|A \cap B|_w$ instead of $|[\![A]\!]_w \cap [\![B]\!]_w|$.

(see the concluding remarks in Section 7 below). Also, the reason we single out the cardinality of the restrictor, rather than of the nuclear scope or both, is because in the examples we focus on, subjects base their judgments primarily on information they receive about the cardinality of the restrictor, making it an essential parameter to how they ascribe *many*.[7] The truth-conditions laid out here agree with those stated by Lappin (2000); in particular they take account of the intensionality of *many* and of the fact that *many* is nonsymmetric with regard to its arguments (*many As are Bs* does not entail *many Bs are As*).[8]

In the experiments we conducted, we asked subjects to evaluate two sentences of the form seen in (9), with B and C forming a pair of predicates with opposite semantic and desirability values.

(9) a. Many As are Bs.
b. Many As are Cs.

That is, B applies to x iff C does not, and B expresses a positively valued outcome, whereas C expresses a negative outcome. Each time, we set a scenario in which the numbers of ABs and ACs are identical, and their proportion to the number of As was $1/2$. As discussed by Partee (1989), sentences of the form *many As are Bs* are generally ambiguous between a cardinal reading ('the number of ABs is greater than some absolute value') and a proportional reading ('the proportion of ABs to As is greater than a threshold'). In all of our scenarios, we expected the proportional reading to be favored, but by setting equal cardinalities and ratios relative to B and C, we could make sure to test the effect of each predicate even in case the cardinal reading might be relevant.

In line with the severity bias, the prediction formulated in Egré 2014 was that subjects would more readily assent to the sentence with a negatively val-

[7] This is also a natural choice to make under the assumption that all natural language quantifiers Q are *conservative*, that is that $Q(A)(B)$ is semantically equivalent to $Q(A)(A \cap B)$ (see Barwise & Cooper 1981). See Greer 2014 for more on the conservativity of *many*.

[8] Lappin's parametric semantics for *many* is given by the following formula, where $N(w)$ denotes the set of normative situations that are relevant relative to the context w:

(8) $[\![many]\!]_w = \lambda P \lambda Q \forall w' \in N(w)(|P \cap Q|_w \geq |P \cap Q|_{w'})$

That is, many As are Bs provided the actual number of ABs is above the number of ABs in each normative alternative to the actual world. Note that (8) provides symmetric truth-conditions for *many As are Bs* and *many Bs are As*, but Lappin retrieves nonsymmetry by imposing appropriate constraints on $N(w)$.

ued predicate as opposed to a positively valued one, thus establishing a shift in the threshold for *many* depending on the predicate under consideration.[9] In particular, like *harm the environment* and *help the environment* in Knobe's scenario, or *getting Down Syndrome* and *getting insomnia* in Pighin's scenario, the first two predicates *die* and *survive* that we tested denote event types with opposite affective values. Moreover, *die* and *survive* are arguably contradictories, assuming *survive* is semantically analyzable as *not die*. Finally, applied to the predicate *children*, they produce a high contrast in expectations.

To compare those predictions to the semantics laid out above for *many*, we designed experiments with similar structures. In each experiment, we first probed for subjects' truth-conditional intuitions in relation to the two target sentences. In a second phase, we asked participants to provide explicit information about the numerical threshold relevant to ascribe *many* for appropriate counterparts of the sentences in the first phase.

3 Experiment 1 (*beaucoup*)

The first study we present in this section was conducted on French-speaking subjects and French text was used in place of the English translations provided here. *Many children died* and *many children survived* in particular translate *beaucoup d'enfants sont morts* and *beaucoup d'enfants ont survécu* respectively (literally, *beaucoup de A* = *many of A*).[10]

3.1 Experiment 1A

3.1.1 Method

In this experiment, we used the following statement:

> 10 children were present in a school when a fire broke out. 5 of the children survived, the other 5 died.

[9] As pointed out concerning *low* versus *high* or *unintentional* versus *intentional*, we note that we do expect the main prediction to be reversed if we had picked *few* or *not many* instead of *many*, that is, the threshold to be lower for the less negative predicate. We did not test that prediction, however, and chose to focus only on *many*, rather than its antonym.

[10] The original French texts for this experiment, as well as the material of Experiments 2–4, are available from the journal's website at http://dx.doi.org/10.3765/sp.8.13s.

50 participants were recruited in the Laboratoire de Sciences Cognitives et Psycholinguistique in Paris. 32 were women and the age mean was 23.8. Half of the participants first were given the following question:

> Would you say that many children survived? ("YES" or "NO")

Then they got a second question:

> Would you say that many children died? ("YES" or "NO")

The other half got the same two questions, but in reverse order (each participant saw the next question only after answering the first). So, we had one variable (the answer) and two factors: the type of *predicate* (*died* or *survived*) and the *order* (first or second).

3.1.2 Results

The percentages of positive answers by condition are summarized in Table 1. To test whether the kind of predicate used in the question had an effect on participants' answers, we used a McNemar's test with continuity corrections. It revealed that the percentage of positive answers significantly differed between the two predicates ($\chi^2(1, N = 50) = 36.03$, $p < .001$). We then used two chi-square tests to test for the presence of order effects. We found no order effect, either for the *died* question ($\chi^2(1, N = 50) = 0.52$, $p = .47$), or for the *survived* question ($\chi^2(1, N = 50) = 0.52$, $p = .47$).

	Died	*Survived*
Answered first	100%	12%
Answered second	92%	28%

Table 1 Percentage of positive answers by condition for Experiment 1A

3.2 Experiment 1B

3.2.1 Method

In this experiment, we used the first part of the statement used in Experiment 1A, namely *10 children were present in a school when a fire broke out*. The subjects were then given the following questions:

i. From which number of children being dead would you say that many children died?

ii. From which number of children having survived would you say that many children survived?

40 participants were recruited in the Laboratoire de Sciences Cognitives et Psycholinguistique in Paris. 34 were women and the age mean was 22.8. Half of participants received both questions in one order and the other half in the reverse order.

3.2.2 Results

As in Experiment 1A, we had one variable (the answer) and two factors: the type of *predicate* (SURVIVED or DIED) and the *order* (FIRST or SECOND). The mean answers by condition are summarized in Table 2. We used a two-factor ANOVA with repeated measures. There was a main effect of *predicate* ($F(1,38) = 51.2$, $p < .001$) but no main effect of *order* ($F(1,38) = 0.1$, $p = .90$) and no interaction effect ($F(1,38) = 0.1$, $p = .70$).

	Died	*Survived*
Answered first	3.5	7.2
Answered second	3.6	7.0

Table 2 Mean answers by condition for Experiment 1B

3.3 Discussion

In Experiment 1A we set equal cardinalities for the number of children dying and of children surviving, namely 5, and we ensured that each group would comprise 1/2 of the children.

The first observation to make about Experiment 1A is that it confirms the prediction at issue; that is, subjects were much more willing to use *many* in relation to the most negatively loaded of the two sentences, despite the fact that the answers to “how many As are Bs?” and to “how many As are Cs?” are the same. From a semantic point of view, the results therefore confirm the fact that *many* does not behave purely as an extensional quantifier: by this we mean that the evaluation of *many As are Bs* is not sensitive merely to the cardinality of *A*s, *B*s or to the ratio of *AB*s to *A*s.

Secondly, we can see only a slight tendency for subjects to be more willing to say that *many children survived* when the question comes second. The lack of significant order effect indicates that subjects are little prone to readjusting the respective threshold they associate with each predicate depending on their previous answer. This observation is more amply confirmed by the results of Experiment 1B, where we also found no significant order effect.

In Experiment 1B, subjects differed from one another in the threshold, between 1 and 10, that they associated with each predicate, consistent with the fact that *many* is a vague quantifier. Few subjects, however, picked identical thresholds for the two predicates *died* and *survived* (4 subjects out of 40), and few set the standard higher for *died* than for *survived* (3 out of 40). That is, most subjects (the remaining 33, or 82.5%) introduced a gap between the two thresholds and selected a lower threshold for the more negatively valued predicate *die*.

Taken together, the data of Experiments 1A and 1B are consistent with the truth-conditions laid out in (7). In particular, if there were many subjects in Experiment 1A for whom, in the fire and school context w under discussion,

$$\mathbf{norm}_w(\mathit{children}, \mathit{die}, |\mathit{children}|_w = 10) < 4$$

and

$$\mathbf{norm}_w(\mathit{children}, \mathit{survive}, |\mathit{children}|_w = 10) \geq 7,$$

then the contrast in truth-values between (9a) and (9b) follows when

$$|\mathit{children} \cap \mathit{die}|_w = |\mathit{children} \cap \mathit{survive}|_w = 5.$$

For those subjects, given the actual number of dead children specified in Experiment 1A, this means that they would have expected more children to survive and fewer children to die.

Experiment 1 is limited in several ways, however. First of all, judgments of *many*-ness are elicited relative to a set of small cardinality: while our results show that a number less than 10 is enough to count as *many* if we pick the right context and predicate, it would be interesting to see whether these asymmetries are affected at all when larger cardinalities are involved. A second and more important worry is that Experiment 1 includes no control of prior expectations regarding the death of children in a school in dramatic circumstances such as a fire, and more generally, it fails to clarify the sense of *expect* that we refer *many* to. Because of that, one may object that we are jumping too quickly to the conclusion that ascriptions of *many* are sensitive to affective or desirability values conveyed by the predicate, as opposed to just an expectation based on probabilities or frequencies for such events. Both limitations are suppressed in Experiment 2.

4 Experiment 2 (*many*)

In this experiment, we used the same two-part protocol as in Experiment 1, and the same pair of predicates (*die* versus *survive*), but this time we introduced two controls. First of all, we used different cardinalities as our backdrop for judgments of *many*-ness, namely 10 (for replication), 60 and 120. Secondly, we controlled for prior expectations of survival and death, to see in particular whether a higher prior expectation of death would imply a decrease in ascriptions of *many*-ness, and a corresponding increase for the use of *survive*. Thirdly, we changed the scenario so as no longer to talk about children and to refer to no specific age class in relation to casualties. Fourthly, we tested anglophone speakers on scenarios expressed directly in terms of the English word *many*.[11]

4.1 Method

In this experiment, we used the two following scenarios:[12]

High Probability “A boat sailing on a shallow river with X passengers on board hits a sandbank and begins to sink. $X/2$ of the X passengers drown and die, $X/2$ are able to reach the shore.”

Low Probability “A boat sailing on the ocean with X passengers on board blows up due to motor malfunction. $X/2$ of the X passengers drown and die, $X/2$ are able to cling to the remains of the boat and, two days later, are saved by a cargo ship that, luckily, sees them.”

[11] Note that we used bare *many* in English, and not the partitive *many of*, but this does parallel the use of *beaucoup* in Experiment 1. In Experiment 1 the French says *beaucoup de*, but without a definite article the preposition *de* ‘of’ does not suffice to make the construction partitive: *beaucoup d’enfants* in French means *many children*. *Many of the children* would be translated by *beaucoup des enfants*.

[12] $X/2$ is shorthand for 5, 30 and 60 respectively when X is 10, 60 and 120 respectively. Participants never saw fractions, but only round numbers.

Our hypothesis was that people would attribute a higher expectation of survival in the High Probability than in the Low Probability scenario. To test this assumption, we gave only the first sentence of both these scenarios to participants recruited on Amazon Mechanical Turk. There were 60 participants, all located in the United States: 20 participants received both scenarios with $X = 10$; 20 others, with $X = 60$; and a third group of 20, with $X = 120$. In each condition, 10 participants were asked "How many passengers do you think will die? (Give an answer between 0 and X)" and 10 were asked "How many passengers do you think will survive? (Give an answer between 0 and X)". To allow for comparison, the participants' answers to the question in each condition, initially given in absolute numbers of passengers, were converted to percentages of passengers. For example, an answer of 6 for $X = 10$ became 60%.[13] Results are presented in Table 3 and validate our hypothesis.[14]

	Died	*Survived*
High Probability	15%	84%
Low Probability	73%	35%

Table 3 Mean estimates (converted to percentages) of the number of passengers who died/survived as a function of scenario (High versus Low Probability) and question asked (*died* versus *survived*)

Then, we gave our scenarios to 240 distinct participants recruited through Amazon Mechanical Turk. Participants were located in United States, age mean was 31.7, and 116 participants were women. Half of the participants received the High Probability case and half received the Low Probability case. In each group, a third of participants received the scenario with X=10; a third, with X=60; and a third, with X=120. Each participant received the four following questions:

[13] Technically, this means that we asked our participants to tell us the expected value, in the statistical sense, of survival versus death for each cardinality. What we present are the underlying average probabilities, as inferred from those values. In the rest of this paper, we often use the two notions interchangeably.

[14] For *died* questions, a two-factor ANOVA revealed a significant effect of condition (High versus Low.) ($F(1,55) = 52.7, p < .001$) but no significant effect of the total number of passengers ($F(1,55) = 0.5, p = .48$). We obtained similar results for *survived* questions, with a significant effect of condition (High versus Low) ($F(1,55) = 32.6, p < .001$) but no significant effect of the total number of passengers ($F(1,55) = 0.8, p = .39$).

i. Would you say that many passengers died? (“YES” or “NO”)

ii. Would you say that many passengers survived? (“YES” or “NO”)

iii. Starting from which number of dead passengers would you agree to say that many passengers died? (Give an answer between 1 and X)

iv. Starting from which number of survivors would you agree to say that many passengers survived? (Give an answer between 1 and X)

In each condition, half of the participants received questions (i) and (iii) first, and then questions (ii) and (iv). The other half received questions (ii) and (iv) first, and then questions (i) and (ii). In this experiment, unlike in Experiment 1, all questions asked were visible simultaneously after the text.

4.2 Results for questions (i) and (ii)

We first present results for questions (i) and (ii). The percentages of positive (“YES”) answers by conditions are summarized in Table 4.

	Died	*Survived*
Answered first	97%	53%
Answered second	81%	68%

Table 4 Percentages of positive answers by question and order for questions (i) and (ii)

Died* versus *Survived In agreement with the results of Experiment 1A, a McNemar’s test with continuity corrections revealed that the percentage of positive answers significantly differed between the two predicates ($\chi^2(1, N = 240) = 49.87$, $p < 10^{-11}$). Participants tended to answer more positively to the *died* than to the *survived* question (see Table 4).

Order Effects For the *died* question, a chi-square test revealed a significant difference between answers given by participants who had read the *died* question first and those given by participants who had read the *survived* question first ($\chi^2(1, N = 240) = 13.6$, $p < .001$). Similarly, for the *survived*

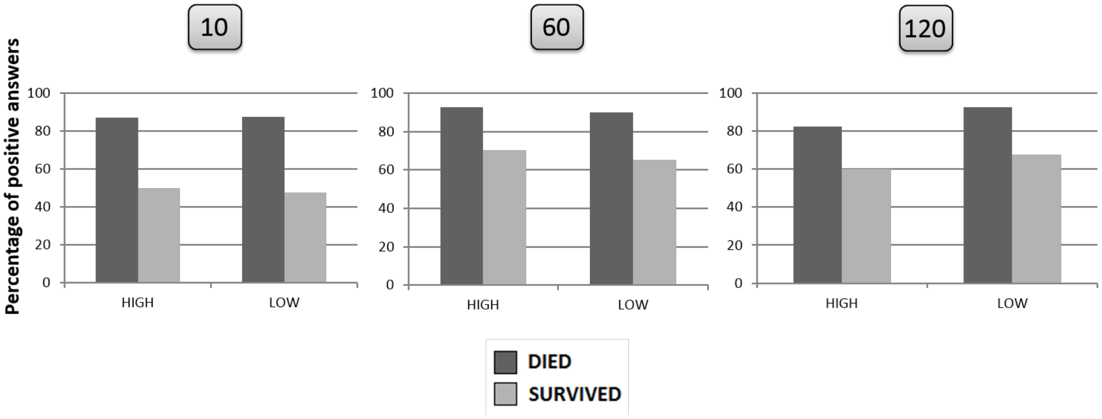

Figure 1 Percentages of positive answers by question, probability of survival, and total number of passengers.

question, a chi-square test revealed a significant difference between answers given by participants who had read the *died* question first and those given by participants who had read the *survived* question first ($\chi^2(1, N = 240) = 9.8$, $p < .01$; see Table 4).

Probability of Survival For the *died* question, a chi-square test revealed no significant difference between participants in the High Probability condition, with 87% giving positive answers, and those in the Low Probability condition, with 90% giving positive answers ($\chi^2(1, N = 120) = 0.2$, $p = .67$). Similarly, for the *survived* question, a chi-square test revealed no significant difference between participants in the High Probability condition, with 61% giving positive answers, and those in the Low Probability condition, with 60% giving positive answers ($\chi^2(1, N = 120) = 0$, $p = 1$; see Figure 1).

Cardinality For the *died* question, a chi-square test revealed no significant effect of cardinality ($\chi^2(2, N = 240) = 0.78$, $p = .68$). However, cardinality had an effect for the *survived* question, with the number of positive answers being significantly higher when the number of passengers was 60 or 120 than when there were only 10 passengers on the boat ($\chi^2(2, N = 240) = 6.6$, $p < .05$; see Figure 1).

4.3 Results for questions (iii) and (iv)

For comparison, the participants' answers to questions (iii) and (iv), initially given in number of passengers, were converted in percentage of passengers. For example, an answer of *4 passengers* to question (iii) for the ten-passenger condition was converted to a value of 40%.

	Died	*Survived*
Answered first	30%	55%
Answered second	40%	55%

Table 5 Mean answers by condition for questions (iii) and (iv)

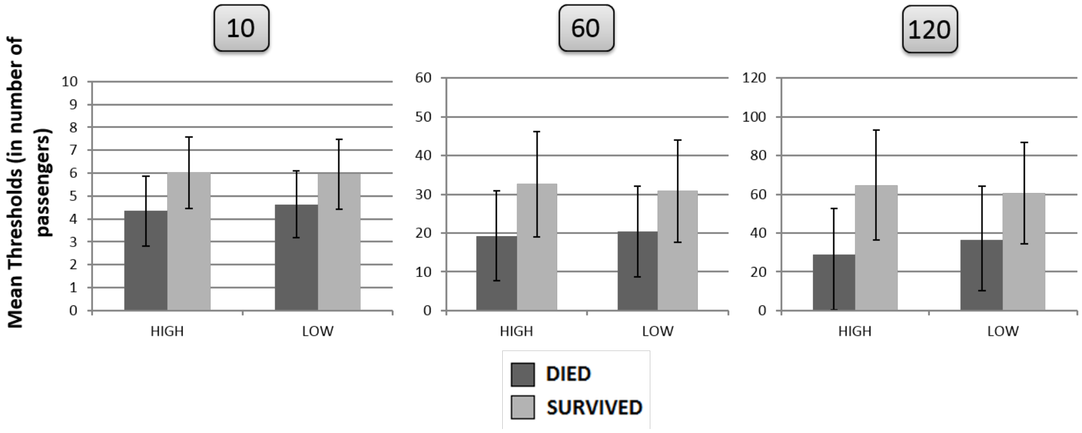

Figure 2 Mean thresholds by question, probability of survival, and total number of passengers. Error bars indicate standard deviation.

***Died* versus *Survived* and Order Effects** We used a two-factor ANOVA with repeated measures. There was a main effect of the question ($F(1,235) = 137.25$, $p < .001$), a main effect of order ($F(1,235) = 7.4$, $p < .01$) and an interaction effect between the two factors ($F(1,235) = 8.7$, $p < .01$; see Table 5).

Probability of Survival For the *died* question, a Welch's t-test revealed no significant difference between participants in the High Probability condition, with a threshold of 33%, and those in the Low Probability condition, with a threshold of 37% ($N = 240$, $t = -1.5$, $df = 233.7$, $p = .15$). Similarly, for the *survived* question a Welch's t-test revealed no significant difference between participants in the High Probability condition, with a threshold of 56%, and those in the Low Probability condition, with a threshold of 54%. ($N = 240$, $t = 0.9$, $df = 233.9$, $p = .35$; see Figure 2).

Cardinality For the *died* question, an ANOVA revealed a significant effect of cardinality ($F(1, 234) = 34.8$, $p < .001$), with thresholds higher in the ten-passenger condition. There was also an effect for the *survived* question ($F(1, 234) = 5.0$, $p < .05$), with thresholds higher in the ten-passenger condition (see Figure 2).

4.4 Discussion

Experiment 2 confirms the main findings of Experiment 1. First of all, we find the same contrast in judgments involving *many* depending on whether the predicate is *die* or *survive*, with higher endorsements for the more negative predicate *die*, irrespective of cardinality and irrespective of the difference in probabilities of survival. Secondly, we find the same correlative contrast in threshold answers — that is, lower threshold values for the more negative predicate. It is worth noting that in the condition involving only 10 passengers, the contrasts are less pronounced than those found in Experiment 1A and 1B — talk of children's life and fire's at school may possibly create more extreme expectations — but they persist nonetheless. Similarly, we see that an increase in cardinality entails a corresponding increase in endorsements of *many* for the *survive* predicate, indicating that *many* is not sensitive merely to proportion in our examples, but indeed to cardinality.

Regarding the main parameters of interest in this experiment, we may have expected higher endorsements of *many* to occur in the high-survival-chance scenarios in comparison to the low-chance ones. Intuitively, the more people one might expect to survive, the more one should say *many* to *die*, and conversely for *survive*. This is not what we observed, however. Prima facie, this suggests that the base probabilities of survival do not essentially affect the judgments concerning *many*, a rather surprising result. This could mean that the probabilities of survival we have elicited are not extreme enough to make such differences appear, or that judgments about *many* for predicates such as *die* and *survive* are essentially based on the consideration of cardinalities and proportions, and not on statistical expectations. Another possibility may be that participants simply ignored that information, unlike participants in the pre-test group. We clarify this issue in what follows.

5 Experiment 3 (*stay* vs. *leave*)

While consistent with each other, the results of the previous two experiments only provide us with a limited basis for the main generalization under discussion, because in both of them we use a single pair of predicates, namely *die* versus *survive*. In order to test the robustness of the asymmetry we found, we ran a third experiment on the model of Experiment 2, but involving a less loaded pair of predicates. In Experiment 3, we used the pair *stay [until the end]* versus *left [before the end]*. Unlike *die* or *survive*, predicates like *stay* and *leave* do not convey any positive or negative value *per se*. In our experiment, a difference in desirability is only *contextually* attached to the *stay*-or-*leave* outcome, in relation to the general attendance at a party. Probabilities, as in Experiment 2, were manipulated via contextual features — in this case, features of the party that should favor leaving early versus staying later.

5.1 Method

In this experiment, we used the following two scenarios:

Successful Party On a university campus, a college fraternity organizes a {small / — / big} student party. X show up to the party. As planned by the organizers, the music at the party is played by a famous and talented band they booked weeks in advance. Also, there is a lot of alcohol, more than students at the party could ever drink. $X/2$ of the students left the party before the end, while $X/2$ others stayed until the end.

Unsuccessful Party On a university campus, a college fraternity organizes a {small / — / big} student party. X show up to the party. However, the famous and talented band that was planned cannot come and is replaced at the last minute by an unknown local band. Moreover, alcohol supplies are too short and run out long before the end of the

> party. $X/2$ of the students left the party before the end, while $X/2$ others stayed until the end.

Our hypothesis was that people would predict that more students would stay and fewer students would leave in the Successful Party than in the Unsuccessful Party scenario.[15] To test this assumption, we gave these scenarios (minus the last sentence indicating the number of students who actually stayed or left) to participants recruited on Amazon Mechanical Turk. There were 60 participants, all located in United States. Each participant received only one scenario (the Successful Party or the Unsuccessful Party). The total number of people showing up to the party (X) varied among participants, with 20 participants receiving cases with $X = 10$; 20 receiving cases with $X = 60$; and 20 receiving cases with $X = 120$. In each condition, participants were asked "How many students do you think stayed until the end of the party? (Give an answer between 0 and X)", and then "How many students do you think left before the end of the party? (Give an answer between 0 and X)". Results are presented in Table 6 and validate our hypothesis.[16]

	Left	*Stayed*
Successful Party	23%	77%
Unsuccessful Party	75%	24%

Table 6 Estimates for the percentage of students who stayed/left in each scenario

We then gave our scenarios to 120 participants recruited through Amazon Mechanical Turk. Participants were located in United States, age mean was 34.5, and 79 participants were women. Half of the participants received the Successful Party case and the other half the Unsuccessful Party case. In each group, a third of the participants received the scenario with $X = 10$; a

[15] The terms "successful" and "unsuccessful" are mere labels for our scenarios, and were never presented to participants.

[16] For *left* questions, a two-factor ANOVA revealed a significant effect of scenario (Successful Party versus Unsuccessful Party) ($F(1,56) = 91.6$, $p < .001$) but no significant effect of the total number of students ($F(1,56) = 0.1$, $p = .82$). We obtained similar results for *stayed* questions, with a significant effect of condition (Successful Party versus Unsuccessful Party) ($F(1,56) = 100.0$, $p < .001$) but no significant effect of the total number of students ($F(1,56) = 0.0$, $p = .97$).

third, with $X = 60$; and a third, with $X = 120$. Each participant received the following six questions:

a. How many students left before the end?

b. How many students stayed until the end?

i. Would you say that many students left before the end? ("YES" or "NO")

ii. Would you say that many students stayed until the end? ("YES" or "NO")

iii. Starting from which number of students leaving before the end would you agree to say that many students left? (Give an answer between 1 and X)

iv. Starting from which number of students staying until the end would you agree to say that many students stayed? (Give an answer between 1 and X)

Questions (a) and (b) were control questions. Participants who failed them were excluded and replaced by newly recruited participants. In each condition, half of participants received first questions (a), (i) and (iii) then questions (b), (ii) and (iv). The other half received them in the reverse order. Between each triplet of questions, participants had to reread the scenario they were assigned.

5.2 Results for questions (i) and (ii)

We first present results for questions (i) and (ii). The percentages of positive ("YES") answers by conditions are summarized in Table 7.

	Left	*Stayed*
Answered first	88%	67%
Answered second	73%	47%

Table 7 Percentages of positive answers by question and order for questions (i) and (ii)

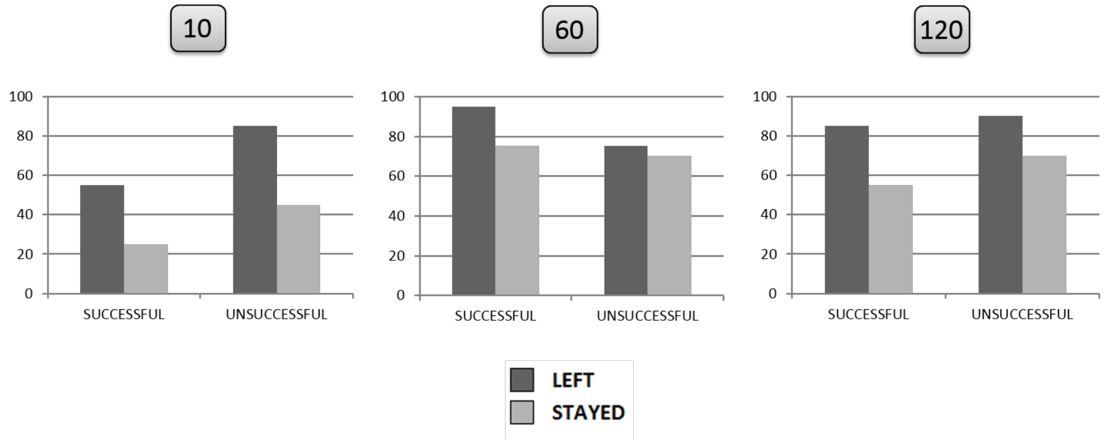

Figure 3 Percentages of positive answers by question, success of the party, and total number of students.

Left* versus *stayed A McNemar’s test revealed a significant difference between answers for the *left* question and answers for the *stayed* question ($\chi^2(1, N = 120) = 12.6$, $p < .001$). Participants tended to answer more positively to the *left* than to the *stayed* question.

Order effects For the *left* question, a chi-square test revealed no significant difference between answers given by participants who had read the *left* question first and those given by participants who had read the *stayed* question first ($\chi^2(1, N = 120) = 3.4$, $p = .06$). For the *stayed* question, a chi-square test revealed a significant difference between answers given by participants who had read the *left* question first and those given by participants who had read the *stayed* question first ($\chi^2(1, N = 120) = 4.1$, $p < .05$).

Probability of staying For the *left* question, a chi-square test revealed no significant difference between participants in the Successful Party condition, with 78% giving positive answers, and those in the Unsuccessful Party condition, with 83% giving positive answers ($\chi^2(1, N = 120) = 0.2$, $p = .64$). Similarly, for the *stayed* question, a chi-square test revealed no significant difference between participants in the Successful Party condition, with 52%

giving positive answers, and those in the Unsuccessful Party condition, with 62% giving positive answers ($\chi^2(1, N = 120) = 0.8$, $p = .36$).

	Left	*Stayed*
10 students	70%	35%
60 students	85%	73%
120 students	88%	63%

Table 8 Percentages of positive answers by question and cardinality for questions (1) and (2)

Cardinality For the *left* question, a chi-square test revealed no significant effect of cardinality ($\chi^2(2, N = 240) = 4.6$, $p = .10$). However, cardinality had an effect for the *stayed* question, with the number of positive answers being significantly higher when the number of students was 60 or 120 than when there were only 10 students showing up to the party ($\chi^2(2, N = 240) = 12.3$, $p < .01$).

5.3 Results for questions (iii) and (iv)

As in Experiment 2, the participants' answers to questions (iii) and (iv), initially given in number of students, were converted to percentages of students.

	Left	*Stayed*
Answered first	42%	51%
Answered second	47%	55%

Table 9 Mean answers by condition for questions (iii) and (iv)

***Left* versus *Stayed* and Order Effects** We used a two-factor ANOVA with repeated measures. There was a main effect of the question ($F(1, 110) = 8.72$, $p < .01$), but no main effect of order ($F(1, 110) = 1.54$, $p = .22$). Overall, thresholds were lower for the *left* than for the *stayed* question.

Probability of Staying For the *left* question, a Welch's t-test revealed no significant difference between participants in the Successful Party condition, with a threshold of 49%, and those in the Unsuccessful Party condition, with a threshold of 48% ($N = 118$, $t = -0.3$, $df = 115.5$, $p = .76$). Similarly, for the *stayed* question, a Welch's t-test revealed no significant difference between participants in the Successful Party condition, with a threshold of 54%, and those in the Unsuccessful Party condition, with a threshold of 52% ($N = 118$, $t = 0.8$, $df = 116.7$, $p = .43$).

	Left	*Stayed*
10 students	52%	61%
60 students	43%	49%
120 students	40%	49%

Table 10 Mean answers by question and cardinality for questions (iii) and (iv)

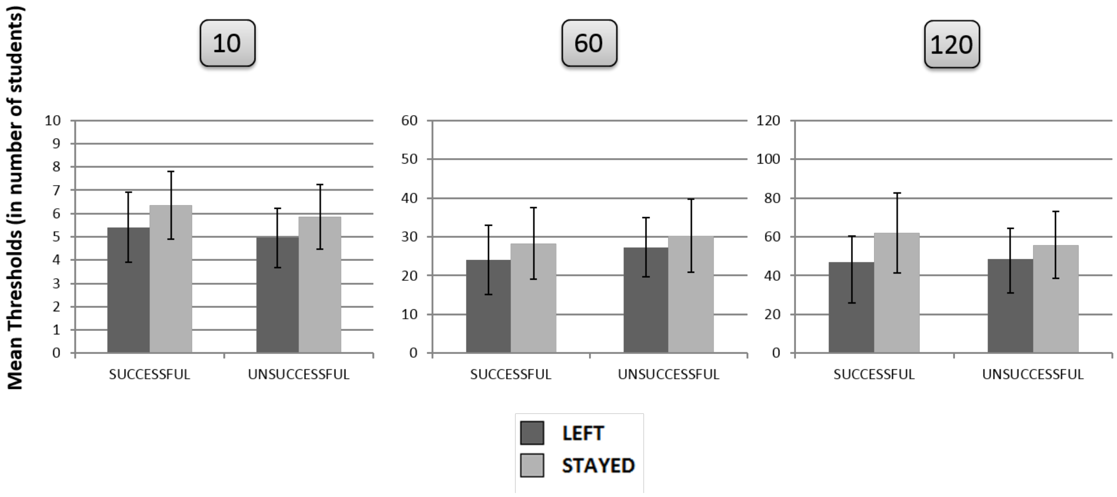

Figure 4 Mean thresholds by question, success of the party, and total number of students. Error bars indicate standard deviation.

Cardinality For the *left* question, an ANOVA revealed a significant effect of cardinality ($F(1, 118) = 15.5$, $p < .001$), with thresholds higher when $N = 10$.

There was also an effect for the *stayed* question ($F(1, 118) = 11.1$, $p < .01$), with thresholds higher when $N = 10$.

5.4 Discussion

This experiment replicates the main findings of Experiment 2. In our scenarios, leaving before the end of the party was meant to be less desirable or positive than staying until the end of the party. As in the previous experiment, categorical ascriptions of *many*-ness are more prevalent for the former than the latter predicate, and similarly, the thresholds attached to leaving are on average lower than those attached to staying. Those differences, moreover, hold similarly across both Successful and Unsuccessful conditions, showing a relative insensitivity to the probabilistic representation of how many might leave or stay in each scenario.

6 Experiment 4 (*pass* vs. *fail*)

In Experiment 2 as well as in Experiment 3, we found no statistically significant effect of the change of probabilistic expectations on judgments involving *many*. Intuitively, however, ascriptions regarding *many* ought to also depend on probabilistic expectations, even when moral expectations regarding the desirability of the outcome appear to prevail. To test for that, we ran a fourth experiment, in which we also used a pair of predicates with analogous status to the ones we used before, this time *pass* versus *fail* in relation to an exam. The main difference with our previous experiments is that we did not leave probabilistic expectations quantitatively implicit. This time, each condition was given with the explicit numerical specification of the base rate of success/failure at the exam.

6.1 Method

We used the following two scenarios:

Easy Exam X students are taking a Chinese exam to prove that they have adequate skills in Chinese. Usually, about 75% of examinees succeed in this exam. This time, $X/2$ of the X students pass the exam, while $X/2$ fail.

Hard Exam X students are taking a Chinese exam to prove that they have adequate skills in Chinese. Usually, about 25% of examinees succeed in this exam. This time, $X/2$ of the X students pass the exam, while $X/2$ fail.

We gave these scenarios to 120 participants recruited through Amazon Mechanical Turk. Participants were located in United States, age mean was 29.9 and 38 participants were women. Half participants received the Easy Exam case and half received the Hard Exam case. In each group, a third of participants received the scenario with X (the total number of students taking the exam) $X = 10$, a third with $X = 60$ and a third with $X = 120$. Each participant received the same six questions asked in Experiment 3, with the predicate *failed the exam* instead of *left before the end* and the predicate *passed the exam* instead of *stayed until the end*. The questions were administered as in Experiment 3.

6.2 Results for questions (i) and (ii)

We first present results for questions (i) and (ii). The percentages of positive ("YES") answers by conditions are summarized in Table 11.

	Failed	*Passed*
Answered first	80%	55%
Answered second	70%	40%

Table 11 Percentages of positive answers by question and order for questions (i) and (ii)

Failed* versus *Passed A McNemar's test revealed a significant difference between answers for the *failed* question and answers for the *passed* question ($\chi^2(1, N = 120) = 20.1$, $p < .001$). Participants tended to answer more positively to the *failed* than to the *passed* question.

Order Effects For the *failed* question, a chi-square test revealed no significant difference between answers given by participants who had read the *failed* question first and those given by participants who had read the *passed*

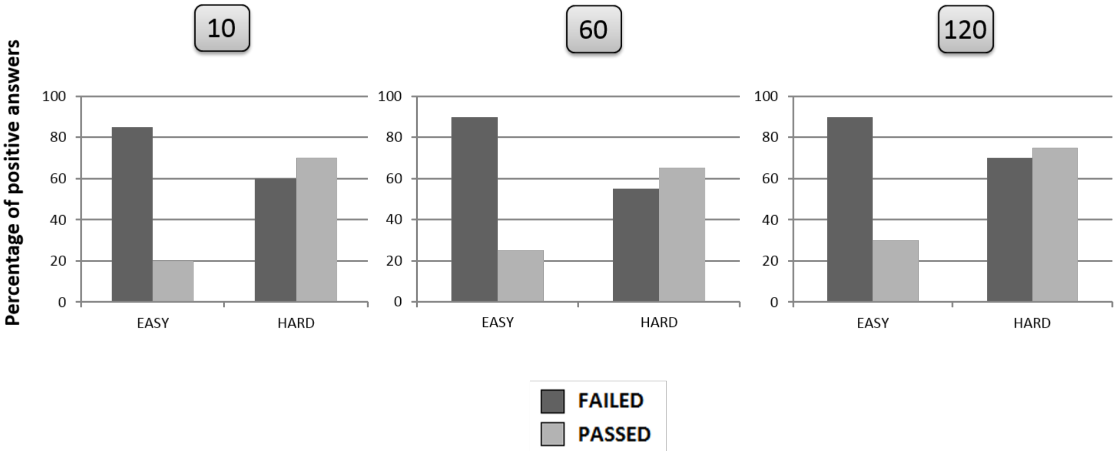

Figure 5 Percentages of positive answers by question, difficulty of the exam, and total number of students.

question first ($\chi^2(1, N = 120) = 1.1$, $p = .29$). For the *passed* question, a chi-square test revealed no significant difference between answers given by participants who had read the *failed* question first those given by participants who had read the *passed* question first ($\chi^2(1, N = 120) = 1.9$, $p = .17$).

Probability of Passing For the *failed* question, a chi-square test revealed a significant difference between participants in the Easy Exam condition, with 88% giving positive answers, and those in the Hard Exam condition, with 62% giving positive answers ($\chi^2(1, N = 120) = 10.0$, $p < .01$). Similarly, for the *passed* question, a chi-square test revealed a significant difference between participants in the Easy Exam condition, with 25% giving positive answers, and those in the Hard Exam condition, with 70% giving positive answers ($\chi^2(1, N = 120) = 23.6$, $p < .001$).

Cardinality For the *failed* question, a chi-square test revealed no significant effect of cardinality ($\chi^2(2, N = 240) = 0.8$, $p = .67$). For the *passed* question, a chi-square test revealed no significant effect of cardinality ($\chi^2(2, N = 240) = 0.5$, $p = .77$).

	Failed	*Passed*
10 students	73%	46%
60 students	73%	45%
120 students	80%	53%

Table 12 Percentages of positive answers by question and cardinality for questions (i) and (ii)

6.3 Results for questions (iii) and (iv)

As in the previous experiments, the participants' answers to questions (iii) and (iv), initially given in number of students, were converted in percentage of students.

	Failed	*Passed*
Answered first	44%	56%
Answered second	49%	55%

Table 13 Mean answers by question and cardinality for questions (iii) and (iv)

***Failed* versus *Passed* and Order Effects** We used a two-factor ANOVA with repeated measures. There was a main effect of the question ($F(1, 118) = 37.4, p < .001$), but no main effect of order ($F(1, 118) = 0.1, p = .73$).

Probability of Success For the *failed* question, a Welch's t-test revealed a significant difference between participants in the Easy Exam condition, with a threshold of 41%, and those in the Hard Exam condition, with a threshold of 52% ($N = 112$, $t = -3.4$, $df = 104.4$, $p = .001$). Similarly, for the *passed* question, a Welch's t-test revealed a significant difference between participants in the Easy Exam condition, with a threshold of 64%, and those in the Hard Exam condition, with a threshold of 47% ($N = 110$, $t = 5.1$, $df = 109.6$, $p < .001$).

	Failed	*Passed*
10 students	49%	57%
60 students	45%	55%
120 students	45%	54%

Table 14 Mean answers by question and cardinality for questions (iii) and (iv)

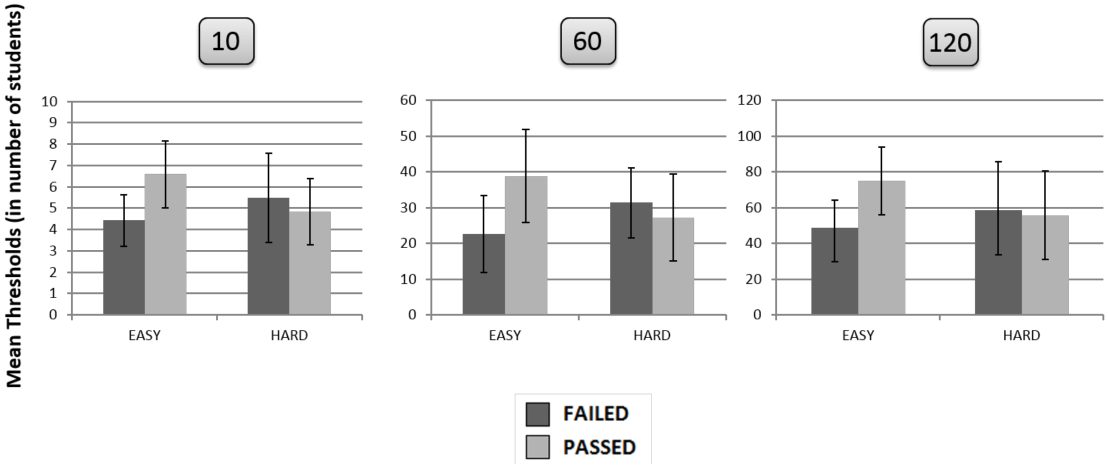

Figure 6 Mean thresholds by question, difficulty of the exam, and total number of students. Error bars indicate standard deviation.

Cardinality For the *failed* question, an ANOVA revealed no significant effect of cardinality ($F(1,112) = 1.2$, $p = .27$). There was also no effect for the *passed* question ($F(1,110) = 0.4$, $p = .55$).

6.4 Discussion

Experiment 4 confirms the main asymmetry found in the previous experiments, namely a differential effect of the predicate related to the most negatively valued event, both in yes-no ascriptions involving *many* and in the selection of a numerical threshold. This time, however, the explicit mention of the base rates does reveal an effect of the probabilistic expectations, both in the yes-no ascriptions and in the setting of a threshold. This shows that both probabilistic expectations and moral expectations are operative in judgments involving *many*, and moreover that they interact.

Consistently with this interaction, the effect of moral expectations still remains dominant. In particular, it would be misguided to infer from our data that the effect of the predicate is operative in the Easy Exam condition and ceases to be operative in the Hard Exam condition. Rather, the data should be interpreted the other way around. If there were no bias toward the most negatively valued predicate in both cases, then the two figures in the Hard Exam and the Easy Exam conditions ought to look like mirror images of each other. In particular, in the Hard Exam condition one might then have expected fewer ascriptions of *many students failed*, knowing that more students usually fail.

The results of Experiment 4 are furthermore consistent with the results obtained in the domain of assessments of probability by Pighin, Bonnefon & Savadori (2011). In the second of their studies on medical risk communication, a variant of the initial scenario was tested in which the doctor informs the patient that the risk communicated is *above average* (in their first study, no relative standard was specified). What they found in this case is that participants become more sensitive to the numerical difference between risk levels in each condition (insomnia versus Down Syndrome), but the severity bias persists nonetheless, with lower probabilities being ranked higher for the less desirable outcome. By specifying an explicit probability standard for success/failure, we thus get an analog of what Pighin calls “overcoming number numbness”: in our case, participants become sensitive to the difference in risk level in their judgments concerning *many*, even though the threshold for *many* remains comparatively lower for the more undesirable outcome.

7 General Discussion

The four experiments presented in this paper confirm several basic observations about the semantics of *many*: that it is context-sensitive, and that judgments of the form *many As are Bs* are not based exclusively on the actual cardinality of the sets *A* and *B*, or on the relative proportion of *AB*s to *A*s. What is new, however, is the evidence concerning the sensitivity of *many* to the affective value or desirability value attached to the predicates under examination. In Experiments 1 and 2, we found that the threshold relevant to ascribe *many* is lower for *die* than for *survive*, in agreement with the truth conditions laid out in section 2, and in a way that comports with the higher rate of assents to *many* for *die* than for *survive*. In Experiment 3 and 4 we found those effects replicated for two other pairs of predicates with a similar structure. Moreover, throughout these experiments, we have observed that these effects were extremely robust: though we sometimes observed order effects suggesting that some participants were willing to make their second answers consistent with the first ones, those were very weak, and presentation of both cases to the same participants never canceled the effect.

A related aspect worth highlighting about our results is the existence of a strong correlation between truth value judgments (yes–no judgments) and judgments about thresholds (numerical judgments). In Experiments 2–4, in which each participant issued both truth value judgments and judgments about thresholds, we can measure the consistency between yes–no answers and threshold answers across participants, by computing the yes–no answer implied by the threshold given for each participant in questions (iii) and (iv), and by comparing it to the actual yes–no answer given in response to questions (i) and (ii). As Table 15 shows, the rate of individual consistency is very high in each condition, with a minimum of 0.8 ($p < 10^{-4}$ in that condition, one-sided binomial test).[17] By themselves, the data do not allow us to conclude that Yes-No answers are computed from the value estimated

[17] In Experiment 1, we can not measure individual consistency, but we can still construct the yes–no answers predicted by participants in Group 1B from their threshold answers, and compare them with those of participants of Group 1A. A Fisher test run for each order in which the question was presented comports with the expectation that we cannot reject the null hypothesis that there is no difference between the two groups ($p = 0.72$ for the *die–survive* order and $p = 0.68$ for the reverse order).

	Predicate	**Cardinality**		
		10	60	120
Experiment 2	*Die*	0.92	0.93	0.94
	Survive	0.93	0.91	0.87
Experiment 3	*Leave*	0.8	0.9	0.9
	Stay	0.8	0.85	0.82
Experiment 4	*Fail*	0.91	0.91	0.94
	Pass	0.94	0.89	0.84

Table 15 Consistency between yes–no answers and threshold answers in Experiments 2–4, for each predicate and cardinality ($p < 10^{-4}$ in all conditions)

by participants as the threshold, but they at least lend support to the view that yes–no answers are derived from the setting of an implicit threshold, as postulated by our model.

In this section we proceed to say more about threshold setting, and we discuss different hypotheses about the way in which probabilistic expectations and moral expectations interact in establishing what number of proportion counts as *many*. We see two main possibilities. The first, following a distinction made by Kahneman & Tversky 1979 between "reference point" and "aspiration level", identifies the standard for *many* with a moral standard (aspiration level), but views the moral standard itself as relative to a statistical norm (reference point). The second possibility is to think of the moral standard and the statistical reference point as being computed independently, letting the threshold for *many* coincide with one or the other depending on the context. We see reasons to favor the first hypothesis, even though more work is needed to adjudicate between the two.

7.1 Moral versus probabilistic expectations

From the literature on *many*, we know that *many* is sensitive to cardinality, to proportion, and to prior expectations concerning those. We see the role of those parameters confirmed, but the main point we take Experiments 2–4 to establish is that this sensitivity to prior expectations cannot be limited to statistical expectations.[18] From Table 3 of Experiment 2, for example, one might have anticipated that in the Low Survival scenarios, where the expectation of death is above 70%, the death of only half of the passengers would impair assents to *many passengers died*, and conversely that the higher proportion of survivors than expected would favor assent to *many passengers survived*. But this is not what we observe. Similarly, in Experiment 4, one might have expected lower endorsements of *many students failed the test*, compared to *many students passed*, in a situation in which it is stipulated that 75% usually fail.

When we say that judgments of many-ness are sensitive to moral considerations that go beyond the representation of statistical facts, we mean to interpret "moral" in the broad sense, not in terms of right and wrong but in terms of good and bad, or desirable and undesirable.[19] As reflection shows, indeed, even as prior probabilities are known or represented, it remains perfectly rational in principle to express such sentences as the following:

(10) a. One might have hoped for more people to survive.
b. One would have liked more people to survive.
c. One might have hoped for fewer people to die.
d. One would have liked fewer people to die.

(And all of the above are compatible with *...although one could expect so many to die*). Likewise, it is perfectly coherent to say things like:

(11) Although I knew the chances, I would have {hoped / preferred} for {more students to pass / fewer students to fail}.

And similarly, it seems perfectly rational to say:

(12) Although one could expect most students to leave before the end of the party, I was hoping {fewer would leave / more would stay}.

A fuller discussion of the contrast between what we call here moral expectations and probabilistic expectations would require a more elaborate investigation of the lexical semantics of attitude verbs such as *hope*, *like* or *prefer*, but lies beyond the scope of this paper. The important point, however,

[18] See in particular Fernando & Kamp 1996 for an account of *many* in which expectations are cashed out exclusively in probabilistic terms.

[19] The term "*evaluative* expectations" may be deemed preferable instead of "*moral* expectations." We hope the intended use of "moral" is sufficiently clear in the present setting.

is that *hope*, *like* or *prefer* all involve an element of desire or wish that *expect* does not have when taken in the purely probabilistic or belief sense of the notion.

From a psychological point of view, our findings fit quite nicely with the idea that the representation of outcomes in sentences of the form *many As are Bs* involves at least two dimensions of comparison: a factual dimension pertaining to what the statistical regularities *are*, and a counterfactual dimension pertaining to what the facts might or could have preferably been. A difficult question concerns the best way to model the interaction between those dimensions. In principle, it would be possible for someone to say:

(13) Although I know that on average only 25% of the students succeed at the Chinese exam, I would have hoped for *all of them* to succeed.

In that case, the speaker should categorically endorse *many students failed* and reject *many students passed*. But then, the same speaker should set the threshold for *many students failed* much lower than what our aggregate data tell us in explicit judgments regarding the setting of the threshold. What the data suggest is that most speakers do not ascribe *many* in relation to such ideal or extreme values. Rather, the data of Experiment 4 suggest that the threshold relevant to ascribe *many* is determined jointly by considerations of the statistical norm and of what may seem more or less desirable *relative to that statistical norm*, rather than relative to what is absolutely or ideally desirable.

Note that the data from Experiments 2 and 3 are not incompatible with this hypothesis. In Experiments 2 and 3, we observe a surprising insensitivity of the judgments to the change in probabilistic expectations. What could be happening is that participants are setting the threshold, depending on the predicate, relative to a common default norm in both conditions. Indeed, in Experiments 2 and 3, the subjects who are tested about *many* are not given an explicit value, nor asked to compute it explicitly (unlike participants in the pre-test group). They may simply focus their attention on the numbers given in each case, and discount the contextual changes to which our pre-test group is sensitive when explicitly instructed to compute those expected values. If so, this could explain why we find an effect of probabilities in Experiment 4 but not in the previous two experiments. Whether or not this is the case, the results of Experiment 4, compared to those of Experiments 2 and 3, suggest to us that probabilistic expectations and moral expectations are not manipulated independently.

7.2 Setting the standard: reference point and aspiration level

The previous remarks invite us to say more about the way in which the standard might be set in ascriptions of the form *many As are Bs*. In (7) we proposed that such sentences are true in a context w iff $|A \cap B|_w \geq \mathbf{norm}_w(A, B, |A|_w)$. As such, the proposed truth conditions encode sufficient information to represent both the cardinal and the proportional reading of sentences involving *many*, and to accommodate extensional as well as intensional readings. They remain silent, however, about the way in which the function **norm** might be computed.

Regarding this point, the data we obtained about *many* bear a connection with empirical evidence at the origin of Kahneman and Tversky's prospect theory (see Kahneman & Tversky 1979). At the bottom of prospect theory is the observation of an asymmetry between losses and gains, namely the observation that *losses loom larger than gains* (Kahneman & Tversky (1979: 279)). In Kahneman and Tversky's framework, a prospect is basically a lottery — that is, a vector of outcomes with various probabilities adding up to 1. The value of a prospect is computed relative to a zero position on a scale for absolute quantities which they call the reference point, with regard to which negative or positive deviations are assigned a utility value. Based on their findings, Kahneman and Tversky derive as one of their main postulates the idea that the value function v associating negative and positive utilities to losses and gains is steeper for losses than for gains.

It would be possible to envisage the results of our experiments in that framework. For example, in Experiment 1, we might consider the reference point to be 0 children dying or surviving, with 5 dead children being seen as a relative loss, and 5 children surviving being seen as a relative gain. The maximal gain, on that perspective, is 10 children surviving, and the maximal loss 10 children dying. One can then represent people's judgments about *many* in relation to Kahneman and Tversky's value function. Firstly, it would mean that $v(5) < -v(-5)$, that is, the positive value of a gain of 5 lives does not compensate the negative value associated with a loss of 5 lives. On top of that, one needs to consider that ascriptions of *many* depend on a common threshold t on the scale of absolute values, such that $v(5) < t < -v(-5)$. That is, 5 children surviving is not many, because it falls below the value threshold, whereas 5 children dying is many, because it falls above the threshold.

If we use that framework, we end up with a unique postulated threshold for *many*, but the computation of *many* first involves representing quantities as gains and losses, and secondly associating a value to them. Instead, our approach uses a single axis to represent cardinalities, but we simply postulate different normative standards for the ascription of *many* along that axis, depending on the predicates. Is the difference between those two approaches substantive?

The first thing to observe in response to this question concerns the relation between the setting of a standard for *many*, and what Kahneman and Tversky call the *reference point*. Kahneman and Tversky's notion of reference point bears some affinity with the notion of standard of comparison in play in the semantics of gradable expressions, but it is not the same notion. Prima facie, Tversky and Kahneman motivate the notion in ways that recall the very example Pettit and Knobe use to account for the Knobe asymmetry about ascriptions of *intentional* to action types. They write:

> When we respond to attributes such as brightness, loudness, or temperature, the past and present context of experience defines an adaptation level, or reference point, and stimuli are perceived in relation to this reference point. Thus, an object at a given temperature may be experienced as hot or cold to the touch depending on the temperature to which one has adapted. The same principle applies to non-sensory attributes such as health, prestige, and wealth. The same level of wealth, for example, may imply abject poverty for one person and great riches for another depending on their current assets.
>
> (Kahneman & Tversky 1979: 277)

This description of the reference point is particularly apt for the account of our data in Experiment 4. In Experiment 4, each scenario specifies an *explicit* reference point in terms of the frequency of success/failure at the exam. We see that judgments concerning *many* are sensitive to the shift of this reference point, for each predicate. For example, one way to conceive the setting of the standard would be to think that if 25% of success fixes the reference point for the predicate *pass* in the Hard Condition, by symmetry (given that $[\![\textit{fail}]\!] = [\![\textit{not pass}]\!]$), the corresponding reference point for *fail* is 75%. In this condition, however, the proportion of students who pass/fail is 50%, meaning that more students pass than what the reference point specifies; on the other hand, fewer fail than what the symmetric point specifies. With regards to (7), assume that $\mathbf{norm}_w(\textit{students}, \textit{fail}, n) = 0.75n$ and $\mathbf{norm}_w(\textit{students}, \textit{pass}, n) = 0.25n$. This would predict that participants with those reference points in mind should respond "yes" to *many students passed* and "no" to *many students failed*. While some participants may behave according to those predictions, this model clearly does not reflect the general shape of our results.

Given our findings, a better model of the setting of the standard may be to identify the normative value postulated in our semantic theory (and by most semantic accounts of gradable expressions) not directly with what Tversky and Kahneman call the reference point, but with what they call an *aspiration level*. According to them, we can think of the reference point as a "status quo position" relative to which deviations are appreciated. But they also acknowledge that "there are situations in which gains and losses are coded relative to an expectation or aspiration level that differs from the status quo" (Kahneman & Tversky 1979: 286). While little is said in their paper regarding this notion of aspiration level, it seems to us to better fit with the outline we gave at the end of the previous section regarding the interplay between probabilistic expectations and moral expectations proper. This notion of aspiration level should also be compared to the account of norms given by Kahneman & Miller 1986.[20] Kahneman and Miller suggest that in many cases, norms are constructed contextually relative to ad hoc counterfactual alternatives, rather than based on knowledge based on general facts or tendencies. This view comports well with our proposal to view the standard as set in relation to what would be a desirable outcome in a given scenario, rather than to what merely happens most of the time or typically.

7.3 Comparative phrases

An alternative possibility is to think of the normative standard of comparison as distinct in principle from both the aspiration level and the reference point, but as potentially coinciding with either of them depending on the case. What could happen is that, depending on the context and predicate at issue, the standard of comparison shifts, so that it is sometimes identifiable with what Kahneman and Tversky call a status quo position (reference point), and sometimes with a counterfactual position (aspiration level).

The availability of overt comparison phrases with *many* lends some support to that hypothesis (see Solt 2011). For example, in the scenario in

[20] We are indebted to an anonymous reviewer on this point.

which 75% of students usually fail the Chinese exam and 50% actually pass, a speaker may judge:

(14) Many students passed *in comparison to how many usually pass.*

But the same speaker may also judge:

(15) Still, many students failed *compared to (just) how many might have otherwise (if things had turned out better).*

For the comparative phrase involving *usually*, "how many students usually pass" would simply pick the average number of students who pass as the standard, or the most frequent number, whereas for the comparative phrase *compared to how many might have [failed]*, the latter would pick some salient counterfactual alternative that the speaker considers to be a more desirable outcome.

Even if things happen in this way, this remains compatible with the idea that the aspiration level is determined relative to some salient status quo position. In other words, the threshold in the salient counterfactual alternative considered by the speaker may still be determined relative to some statistical norm, rather than independently. Further work needs to be done to test this hypothesis. Whether or not this is the case, our results at least rule out the systematic identification of the normative threshold with the reference point, understood as contributing mere statistical information.

As our final point for discussion, we note that overt comparative phrases with *many* may give us further insight into the asymmetries we found, and into the understanding of antonym pairs like *die/survive* in relation to *many*. Consider the following sentence:

(16) Many children died in comparison to how many might have died (if things had turned a little better).

Suppose as in Experiment 1 that the actual number $|A|_w$ of children is 10, the number of children having died is 5 ($|A \cap C|_w = 5$), and let the sentence *how many might have died* pick some preferred counterfactual alternative the speaker has in mind, in which only 3 children die out of 10, and 7 survive. *Many children died [in comparison to how many might have]* would be true provided that the ratio

$$\frac{|A \cap C|_w}{|A \cap C|_{w'}} = \frac{5}{3}$$

is high enough. Conversely, *many children survived [compared to how many might have]* would turn out false if the ratio

$$\frac{|A \cap \overline{C}|_w}{|A \cap \overline{C}|_{w'}} = \frac{5}{7}$$

is not high enough.[21]

Obviously both judgments are compatible, in particular if *high enough* denotes a common threshold of $\frac{1}{1}$ in both cases. The choice of this threshold may be motivated as follows. Our analysis of *many* in (7) states that *many As are Bs* is true in w iff $|A \cap B|_w \geq \mathbf{norm}_w(A, B, |A|_w)$. Except when $\mathbf{norm}_w(A, B, |A|_w)$ is equal to 0, (7) is equivalent to requiring

$$\frac{|A \cap B|_w}{\mathbf{norm}_w(A, B, |A|_w)} \geq 1.$$

Now, let $\mathbf{norm}_w(A, B, |A|_w) = |A \cap B|_{w'}$, for some salient alternative w', and assume furthermore that $\mathbf{norm}_w(A, \overline{B}, |A|_w) = |A \cap \overline{B}|_{w'}$ for the same w'. Let $|A \cap B|_w = |A \cap \overline{B}|_w$, as in our examples (as many children died as children survived), and also $[\![A]\!]_w = [\![A]\!]_{w'}$ (the children are the same in the actual world and in the counterfactual world w'). Under those assumptions, it can be proved that if

$$\frac{|A \cap B|_w}{|A \cap B|_{w'}} \geq 1,$$

then

$$\frac{|A \cap \overline{B}|_w}{|A \cap \overline{B}|_{w'}} \leq 1.$$

This follows from the fact that $|A \cap B|_{w'}$ and $|A \cap \overline{B}|_{w'}$ must add up to $|A|_w$ in this case.

[21] The present account should be compared to Greer's recent treatment of *many*. According to Greer 2014, *many As are Bs* is true in a context iff the ratio of $|A \cap B|$ to $|A \cap U|$ is above a certain threshold (assumed to lie between 0 and 1, in her account), where U is a contextual domain restriction. Greer argues that her account of *many* is fundamentally extensional, as opposed to intensional (in particular, unlike us, she does not relativize extensions to indices). We find a lot of appeal in the idea of handling *many* in terms of proportions, as well as in other aspects of Greer's theory, but we remain of the opinion that *many* is intensional, as soon as comparative phrases can introduce the consideration of counterfactual alternatives. In this, our account is closer to Solt (2011: 163), according to whom "in addition to comparison classes over individuals, sets of individuals and times, we also need comparison classes over worlds". We see further reasons to disagree with Greer on the idea that her own account is extensional, even as restricted to the examples she considers, but we leave a more detailed discussion of this point for another occasion.

Concretely, the latter assumption amounts to saying that the threshold for *many As* in relation to a predicate B could be inferred from the threshold applicable to the antonym of B, and conversely, assuming a fixed perspective. It would therefore predict that participants tend to deny *many As are C* when they assent to *many As are Bs* and when C is the antonym of B, except possibly when the corresponding ratios are 1 for each sentence, and unless they should change their comparison point when going from B to C.

Looking back at our data, we can observe that besides being generally consistent between their answers to questions (i–ii) and (iii-iv) (see again Table 15), a non-negligible number of participants pick thresholds for (iii) and (iv) that add up to *exactly* the cardinality of the restrictor (the proportions are of 30%, 46.7% and 36% respectively in Experiments 2–4). One way to interpret this is simply to think that in response to questions (iii) and (iv), participants rely on an approximation to the threshold they use for actual ascriptions of *many*, rather than on an exact value (as already stressed, we cannot expect participants to pick the exact same value in the two tasks). A confirmation of this may be found in the observation that the average of the sum of the thresholds in each cardinality in Experiments 2–4 never lies too far from the cardinality of the restrictor. An alternative possibility is that thresholds for antonyms really are manipulated independently, as our basic semantics in (7) allows for. Here too, more work is needed to adjudicate between those competing possibilities.

8 Concluding remarks

In this paper we have sought to establish that at least some judgments involving *many* are sensitive to moral expectations that differ from expectations based purely on an estimate of chances, but that take into account an element of desirability. By way of qualification, note that what counts as desirable or undesirable can obviously vary enormously depending on the context and the goals under discussion. For instance, if a group of 10 children were presented as an unusual crowd of 10 organized and relentless murderers, one might imagine that 5 of them dying does not count as *many children died*, and possibly counts as *many children survived*.

Such observations, however, only show that the desirability or undesirability attached to an outcome does not intrinsically depend on the predicate used, but on features of the context. In most of our examples, we could rely on what we may call default moral expectations, namely expectations typically shared by a community on what counts as undesirable or undesirable in a given context. For example, as pointed out by B. Spector (p.c.), the contrast we found between *many children died* and *many children survived* should be linked with the one we can feel between:

(17) a. #Only 5 children died (out of 10).
b. Only 5 children survived (out of 10).

Only 5 As are Bs can only be used if the speaker expected more As to be B than actually happened. Here (17)-a is marked because it can only be uttered by someone who expected more children to die, against the default moral norm.[22] Geurts 2009 discusses similar contrasts based on the operator *it is good that*, which can be adapted to the same example:

(18) a. #It is good that 5 children died (out of 10).
b. It is good that 5 children survived (out of 10).

Here again, (18)-a is the marked case, since uttering it implies that, had more than 5 children died, it would still have been good (see Sanford, Dawydiak & Moxey 2007, Geurts 2009 and Nouwen 2011 for more on monotonicity constraints in relation to framing effects). Or consider the following pair, originally presented in Zuber 1986 to illustrate a different point:

(19) a. # Bill regrets that the glass is half-full.
b. Bill regrets that the glass is half-empty.

(19)-a is marked here since presumably, X *regrets that* $P(y)$ implies that X would have liked y not to be at least as much P. Hence (19)-a implies that Bill would have liked the glass to be less than full, which goes against the default desirability expectation in a context in which no specific information is given about what kind of liquid is in the glass.

The asymmetry we found in moral judgments about *many* thus appears to bear more than a family resemblance with the asymmetries uncovered by Knobe in judgments about intentional action. Each time, opposite judgments can be derived from a shift in moral expectations that depends on the property of which the predicate is predicated (*help* vs *harm* in Knobe's scenarios, *survive* vs *die* and analogous pairs in our scenarios). This connection is valuable, since it sets the Knobe effect on a continuum with so-called framing

[22] See Sapir 1944 for observations about *only* and other adverbial expressions that emphasize the interplay between grading and "affect" in the interpretation of number sentences.

effects, and it shows that both kinds of effect are susceptible of semantic analysis using familiar semantic tools.

References

Bartsch, Renate & Theo Vennemann. 1972. The grammar of relative adjectives and comparison. *Linguistische Berichte* 20. 19–32. http://dx.doi.org/10.1007/3-540-07016-8_11.

Barwise, Jon & Robin Cooper. 1981. Generalized quantifiers in natural language. *Linguistics and Philosophy* 4. 159–219. http://dx.doi.org/10.1007/BF00350139.

Bonnefon, Jean-François & Gaëlle Villejoubert. 2006. Tactful or doubtful?: Expectations of politeness explain the severity bias in the interpretation of probability phrases. *Psychological Science* 17. 747–751. http://dx.doi.org/10.1111/j.1467-9280.2006.01776.x.

Cova, Florian, Emmanuel Dupoux & Pierre Jacob. 2012. On doing things intentionally. *Mind and Language* 27(4). 378–409. http://dx.doi.org/10.1111/j.1468-0017.2012.01449.x.

Egré, Paul. 2010. Qualitative judgments, quantitative judgments and norm-sensitivity. *Behavioral and Brain Sciences* 33(4). 335–336. http://dx.doi.org/10.1111/j.1468-0017.2012.01449.x.

Egré, Paul. 2014. Intentional action and the semantics of gradable expressions: (on the Knobe effect). In B. Copley & F. Martin (eds.), *Causation in grammatical structures*, 176–205. (First version of the paper released 2010). Oxford, UK: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199672073.001.0001.

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

Fernando, Tim & Hans Kamp. 1996. Expecting many. *Semantics and Linguistic Theory (SALT)* 6. 53–68.

Geurts, Bart. 2009. Goodness. *Amsterdam Colloquium* 17. M. Aloni, H. Bastiaanse, T. De Jager, van Ormondt P. & K. Schulz (eds.). 277–285.

Greer, Kristen A. 2014. Extensionality in natural language quantification: The case of *many* and *few*. *Linguistics and Philosophy*. http://dx.doi.org/0.1007/s10988-014-9157-5.

Kahneman, Daniel & Dale T. Miller. 1986. Norm theory: Comparing reality to its alternatives. *Psychological Review* 93(2). 136–153. http://dx.doi.org/10.1037/0033-295X.93.2.136.

Kahneman, Daniel & Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47(2). 263–292. http://dx.doi.org/10.2307/1914185.

Keenan, Edward L. & Jonathan Stavi. 1986. A semantic characterization of natural language determiners. *Linguistics and Philosophy* 9. 253–326. http://dx.doi.org/10.1007/BF00630273.

Kennedy, Chris. 2007. Vagueness and grammar: The semantics of absolute and relative gradable adjectives. *Linguistics and Philosophy* 30. 1–45. http://dx.doi.org/10.1007/s10988-006-9008-0.

Knobe, Joshua. 2003. Intentional action and side-effects in ordinary language. *Analysis* 63. 190–193. http://dx.doi.org/10.1111/1467-8284.00419.

Krumpal, Ivar, Heiko Rauhut, Dorothea Böhr & Elias Naumann. 2011. The framing of risks and the communication of subjective probabilities for victimizations. *Quality and Quantity* 45. 1331–1348. http://dx.doi.org/10.1007/s11135-010-9336-6.

Lappin, Shalom. 2000. An intensional parametric semantics for vague quantifiers. *Linguistics and Philosophy* 23. 599–620. http://dx.doi.org/10.1023/A:1005638918877.

Mandelbaum, Eric & David Ripley. 2010. Expectations and morality: A dilemma. *Behavioral and Brain Sciences* 33(4). 346. http://dx.doi.org/10.1017/S0140525X10001822.

Moxey, Linda M. & Anthony J. Sanford. 1993. Prior expectation and the interpretation of natural language quantifiers. *European Journal of Cognitive Psychology* 5(1). 73–91. http://dx.doi.org/10.1080/09541449308406515.

Moxey, Linda M. & Anthony J. Sanford. 2000. Communicating quantities: A review of psycholinguistic evidence of how expressions determine perspectives. *Applied Cognitive Psychology* 14. 237–255. http://dx.doi.org/10.1002/(SICI)1099-0720(200005/06)14:33.0.CO;2-R.

Nouwen, Rick. 2011. Degree modifiers and monotonicity. In P. Egré & N. Klinedinst (eds.), *Vagueness and language use*, 146–164. London, UK: Palgrave Macmillan.

Partee, Barbara. 1989. Many quantifiers. *Eastern States Conference on Linguistics* 5. 383–402.

Pettit, Dean & Joshua Knobe. 2009. The pervasive impact of moral judgments. *Mind and Language* 24(5). 586–604. http://dx.doi.org/10.1111/j.1468-0017.2009.01375.x.

Pighin, Stefania, Jean-François Bonnefon & Lucia Savadori. 2011. Overcoming number numbness in prenatal risk communication. *Prenatal Diagnosis* 31(8). 809–813. http://dx.doi.org/10.1002/pd.2771.

Sanford, Anthony J., Eugene J. Dawydiak & Linda M. Moxey. 2007. A unified account of quantifier perspective effects in discourse. *Discourse Processes* 44. 1–32. http://dx.doi.org/10.1080/01638530701285556.

Sapir, Edward. 1944. Grading: A study in semantics. *Philosophy of Science* 11(2). 93–116.

Solt, Stephanie. 2009. *The semantics of adjectives of quantity*. New York, NY.

Solt, Stephanie. 2011. Vagueness in quantity: Two case studies from a linguistic perspective. In P. Cintula, C. Fermueller, L. Godo & P. Hajek (eds.), *Understanding vagueness: Logical, philosophical, and linguistic perspectives*, 157–174. London, UK: College Publications.

Tversky, Amos & Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science* 211(4481). 453–458. http://dx.doi.org/10.1126/science.7455683.

Weber, Elke U. & Denis Hilton. 1990. Contextual effects in the interpretation of probability words: Perceived base rate and severity of events. *Journal of Experimental Psychology: Human Perception and Performance* 16(4). 781–789.

Zuber, Ryszard. 1986. Semantic restrictions on certain complementizers. *International Congress of Linguists* 13. 434–436.

Paul Egré
Institut Jean-Nicod (CNRS-ENS-EHESS)
École Normale Supérieure
Département d'Etudes Cognitives
29, rue d'Ulm
75005 - Paris - France
paul.egre@ens.fr

Florian Cova
Swiss Centre for Affective Sciences
University of Geneva
Campus Biotech, CISA - Case Postale 60
CH - 1211 Genève 20 - Switzerland
florian.cova@gmail.com