Meaning and use in the expression of estimative probability

Abstract

Words of estimative probability (WEPs), such as 'possible' and 'a good chance', provide an efficient means for expressing probability under uncertainty. Current semantic theories assume that WEPs denote crisp thresholds on the probability scale, but experimental data indicate that their use is characterised by gradience and focality. Here, we implement and compare computational models of the use of WEPs to explain novel production data. We find that, among models incorporating cognitive limitations and assumptions about goal-directed speech, a model that implements a threshold-based semantics explains the data equally well as a model that semantically encodes patterns of gradience and focality. We further validate the model by distinguishing between participants with more or fewer autistic traits, as measured with the Autism Spectrum Quotient test. These traits include communicative difficulties. We show that these difficulties are reflected in the rationality parameter of the model, which modulates the probability that the speaker selects the pragmatically optimal message.

Keywords: probability, language, pragmatics, semantics, computational model

Word count: 4,156

Meaning and use in the expression of estimative probability

Our ability to express probability is of great importance in daily and scientific life. Sometimes, we can use precise numbers when referring to probabilities; for example, we might say that the probability of a fair coin landing on heads is 50%. But very often, we do not—or cannot—know the exact probability of a particular event. In those cases, we might prefer to use what Kent (1964) called *words of estimative probability* (WEPs) to provide a vague estimate of the actual probability (Erev & Cohen, 1990; Juanchich & Sirota, 2019). The class of WEPs is highly diverse, ranging from simple words (e.g., 'possible', 'likely') to complex phrases (e.g., 'more often than not', 'a small but real possibility').

Because of their central importance, the meaning and use of WEPs has been studied extensively across many disciplinary boundaries (e.g., Beyth-Marom, 1982; Friedman & Zeckhauser, 2015; Kratzer, 1991; Shinagare et al., 2019; Wallsten, Budescu, Rapoport, Zwick, & Forsyth, 1986). From this line of research, two radically different views on the meanings of WEPs have emerged.[1]

The first view holds that sentences containing WEPs have crisp *truth conditions* (e.g., Kratzer, 1991). According to this view, sentences with WEPs carve up the space of possibilities into those where the sentence is true and those where it is false. More specifically, many current proposals argue that WEPs denote *thresholds* on the probability scale (e.g., Lassiter, 2019; Moss, 2015; Swanson, 2006; Yalcin, 2007). Thus, the meanings of 'a good chance', 'possible', and 'unlikely' can be defined as follows, where 'P(x)' stands for the probability of an event x:

(1) a. $[\![\text{there is a good chance that } x]\!] = [\text{P}(x) > \text{P}(\text{not-}x)]$

[1] A note on terminology: we use the term 'meaning' to narrowly refer to the conventional content of an expression rather than what someone who uses that expression conveys, i.e., to *semantic* rather than *pragmatic* meaning.

b. $[\![\text{it is possible that } x]\!] = [\text{P}(x) > 0]$

c. $[\![\text{it is unlikely that } x]\!] = [\text{P}(x) < \text{P}(\text{not-}x)]$

An alternative view holds that the meanings of WEPs are gradient and centered around small areas of prototypical use (e.g., Bocklisch, Bocklisch, & Krems, 2012; Jaffe-Katz, Budescu, & Wallsten, 1989; Zimmer, 1983). This *prototype-based* approach is often couched within the framework of *fuzzy logic*. Whereas the truth-conditional view assumes that sentences with WEPs are always either true or false, fuzzy logic argues that they can be true or false to varying degrees (e.g., Zadeh, 1983, 1996).

To illustrate the contrast between the two views, Fig. 1 visualises hypothetical threshold-based and prototype-based meanings for three WEPs.

Apparent support for the prototype-based approach comes from experimental data on the *use* of WEPs. Invariably, such data show that people associate WEPs with gradient and focalised ranges on the probability scale (e.g., Lichtenstein & Newman, 1967; Reagan, Mosteller, & Youtz, 1989; Willems, Albers, & Smeets, 2020). For example, Mosteller and Youtz (1990) report that their participants associated 'possible' with a median probability of 38.5% and an interquartile range of 42.7%, which suggests that 'It is possible that x' implies that the probability of x lies between 17% and 60%, but most likely around 40%.

According to the prototype-based approach, such patterns of gradience and focality must be reflected in the underlying semantics of WEPs. Indeed, the prototype-based approach essentially proposes that meaning *is* use (Budescu & Wallsten, 1995; Clark, 1990). By contrast, the truth-conditional approach is *modular* in that it takes meaning to be an independent level of representation that requires a separate pragmatic module to connect to actual language use (Partee, 1999, 2001). An important challenge for the truth-conditional approach is to flesh

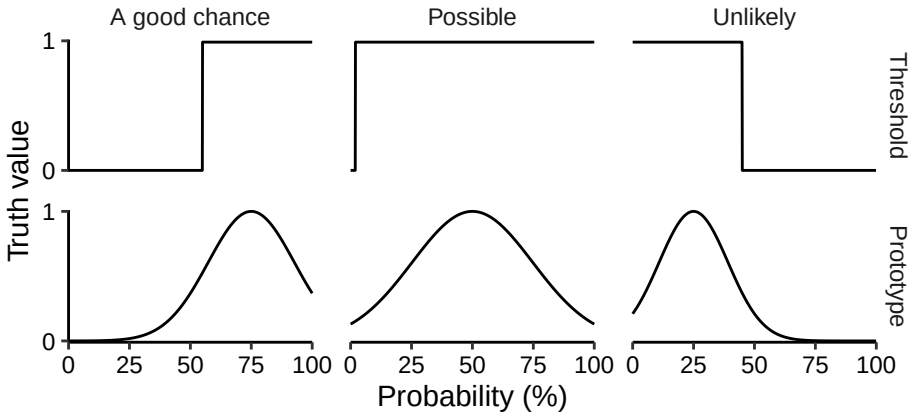

Figure 1: Example threshold-based and prototype-based meanings for three WEPs.

out this pragmatic module to demonstrate how its sparse meanings give rise to complex patterns in language use.

In this paper, we show how a threshold-based semantics for WEPs gives rise to patterns of gradience and focality in their use, once the semantics is embedded within a pragmatic framework that models speaker behaviour as boundedly *rational* (cf. Frank & Goodman, 2012; Franke, 2009; Goodman & Frank, 2016). Here, 'rational' means that speakers prefer to produce those WEPs that are the most likely to receive the intended interpretation on the part of the hearer (cf. Grice, 1975). We show that a threshold-based approach that is embedded in such a model of pragmatic communication offers an equally compelling account of novel data on the production of WEPs as a prototype-based approach that directly encodes gradience and focality into the meanings of WEPs.

Our study builds upon an earlier study by van Tiel, Franke, and Sauerland (2021). In that study, it was shown that patterns of gradience and focality in the use of quantity words (e.g., 'some', 'all') could be reconciled with a truth-conditional

view on their underlying semantics. Here, we examine whether that conclusion generalises from the quantity domain to the domain of probability.

A secondary goal of this study is to validate probabilistic pragmatic models by comparing the production behaviour of people with a low and high *autism spectrum quotient* (AQ) (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001). AQ is a quantitative measure of the extent to which individuals exhibit traits that are associated with *autism spectrum disorder* (ASD). These traits include difficulties with pragmatic communication (American Psychiatric Association, 2013), though the source and scope of these difficulties have been a matter of intense debate (e.g., Baron-Cohen, 1995; Chevallier, Kohls, Troiani, Brodkin, & Schultz, 2012; Kissine, 2012, 2021). We investigate whether such self-reported pragmatic difficulties are reflected in the model parameters; specifically, in a rationality parameter that modulates the probability with which the speaker selects the pragmatically optimal message.

The next section describes our production experiment. Afterwards, we describe the computational model which we use to answer our two research questions, viz. (i) whether patterns of gradience and focality in the use of WEPs can be explained on the basis of a threshold-based semantics, and (ii) whether participants with more autistic traits are less likely to select the pragmatically optimal message than participants with fewer autistic traits.

Production

With some exceptions (e.g., Herbstritt & Franke, 2019; Karelitz & Budescu, 2004; Schuster & Degen, 2020), previous experimental studies have investigated how hearers *interpret* WEPs (e.g., Alstott & Jasbi, 2020; Elsaesser & Henrion, 1990; Renooij & Witteman, 1999); by contrast, in this study, we investigate how speakers naturally *produce* WEPs.

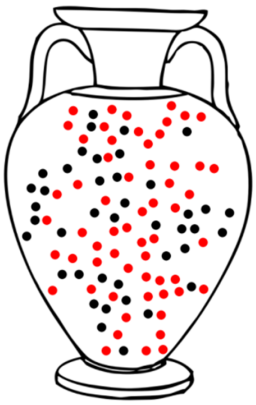

Figure 2: Example display used in the production experiment.

An important advantage of measuring production behaviour is that experiments that measure the interpretation of WEPs typically require participants to engage in *metalinguistic reasoning*. For example, Mosteller and Youtz (1990) asked participants to associate WEPs with ranges on the probability scale. This task requires participants to actively reflect on the meanings of WEPs, and might cause them to draw potentially artificial semantic distinctions between the WEPs that are presented throughout the experiment. As a consequence, it is unclear whether the experimental task measures meaning, use, or participants' beliefs about meaning or use, which problematises the interpretation of the data. Here, we ask participants to describe displays, which is intuitively less likely to invite active reasoning about the meaning and use of WEPs.

For our production experiment (Exp. 1), we recruited 255 participants on Mechanical Turk.[2] Participants were presented with displays showing vases containing 100 randomly distributed marbles (e.g., Fig. 2). The marbles were either red or black. One display was created for each of the 101 possible distributions

[2]We refer the reader to the Appendix for more details about the production experiment. All data and analysis files are available at [removed for anonymous review].

of black and red marbles, and each participant saw a random selection of 25 displays. Participants were asked to describe these displays by freely completing the sentence frame 'If you randomly take a marble from this vase, _____ that it is red'. Participants were instructed not to use numbers or percentages.

We used a relatively open-ended sentence frame rather than one that steered participants towards using WEPs from a specific part of speech. Our motivation for this decision was that we wanted to see which WEPs naturally come to mind, and to accommodate different response preferences observed in previous studies (Budescu, Weinberg, & Wallsten, 1988; Karelitz & Budescu, 2004).

In total, participants produced 1,379 unique responses. Here, we analyse only the 24 WEPs that were mentioned at least 50 times, plus the prominent boundary WEPs 'impossible' and 'certain'. This selection consists of 15 adjectival and 11 nominal WEPs. Similarly to previous studies, we observed distinct response patterns: of the 185 participants who produced at least five WEPs that were included in the analysis, 50 produced exclusively adjectival WEPs; 26 only nominal ones. The remaining 109 participants produced both adjectival and nominal WEPs, suggesting that most participants naturally use a mix of both types of expressions.

Fig. 3 shows the production probabilities of the WEPs in our sample. The results clearly show that participants associate WEPs with gradient and focalised ranges on the probability scale. Can these patterns of use be reconciled with the truth-conditional idea that sentences with WEPs are always either true or false? Or do they necessitate the incorporation of gradience and focality into the semantics of WEPs, as argued for by the prototype-based approach?

To answer these questions, we make use of the computational model of language use introduced by van Tiel et al. (2021), which in turn is based upon the more general Rational Speech Act framework (e.g., Frank & Goodman, 2012; Goodman & Frank, 2016). Here, we give a brief overview of the model.

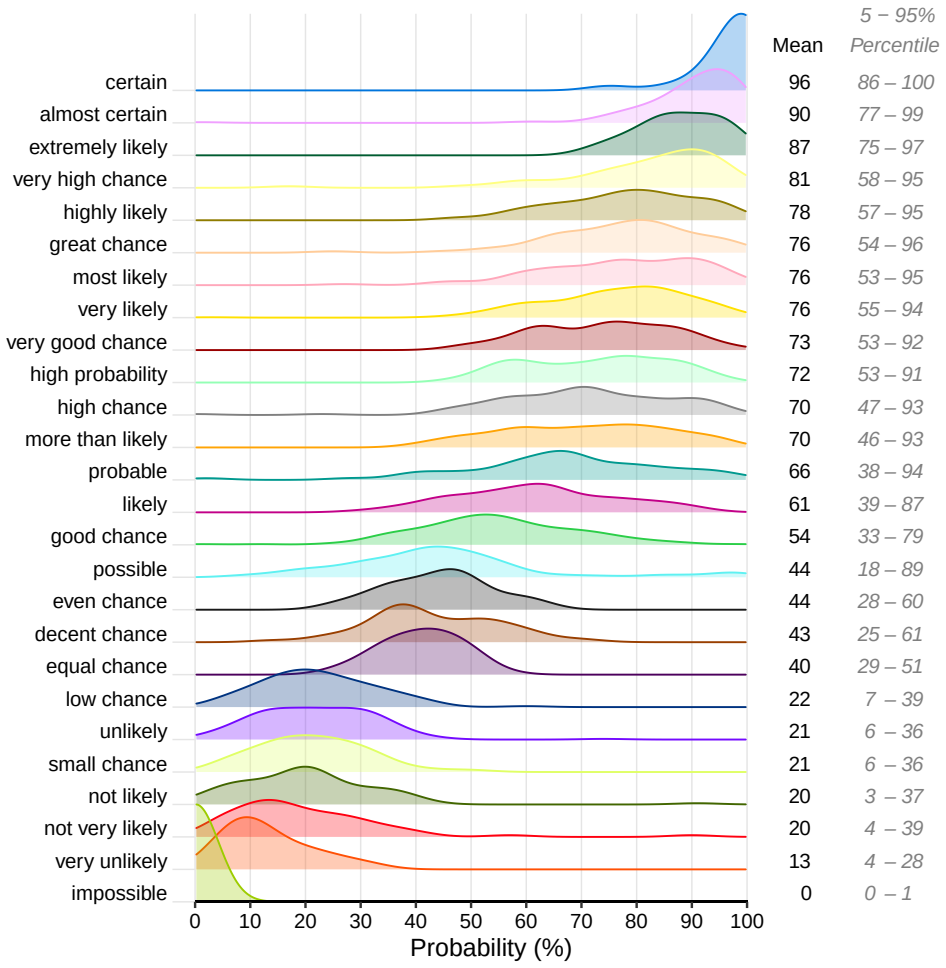

Figure 3: Density plot showing the production probabilities of the most frequently produced WEPs in the production experiment. Next to the plot are the mean probabilities in which WEPs were produced and the corresponding 5–95 percentiles.

Model

The model takes as its point of departure a *lexicon* that associates each pair of a message $m \in M$ and a state of affairs $t \in T$ with a truth value. The set of messages consists of the 26 WEPs in our sample, i.e., $M = \{m_{\text{almost certain}}, m_{\text{certain}}, ...\}$. The set of states consists of the 101 possible probabilities of drawing a red marble. In our design, these probabilities corresponded to the number of red marbles, i.e., $T = \{t_0, t_1, ..., t_{100}\}$.

We define and compare two types of lexica: a threshold-based lexicon that associates each WEP with a threshold on the probability scale, and a prototype-based lexicon that associates each WEP with a gradient and focalised range.

Threshold-based lexicon

The threshold-based approach argues that WEPs denote thresholds on the probability scale. The type of threshold associated with a WEP depends on its *monotonicity*, i.e., its inferential potential. Monotone increasing WEPs like 'possible' license inferences from sets to supersets, e.g., from ordering salmon to ordering fish, as shown by the validity of the argument in (2). By contrast, monotone decreasing WEPs like 'impossible' license inferences from sets to subsets, as shown in (3).

(2) It is possible that he ordered salmon.
→ It is possible that he ordered fish.

(3) It is impossible that he ordered fish.
→ It is impossible that he ordered salmon.

Monotone increasing WEPs place a lower bound on the probability scale; monotone decreasing ones an upper bound.

We determined the monotonicity of the WEPs in our sample by consulting our intuitions about the validity of arguments such as (2) and (3). Consequently, the following WEPs were classified as monotone decreasing: 'not likely', 'not very likely', 'unlikely', 'very unlikely', and 'impossible'. All other WEPs were classified as monotone increasing.

Based on the foregoing, we may define a threshold-based lexicon $\mathfrak{L}_{\text{TH}}$. This lexicon associates each message m with a threshold θ, so that the truth value of m in state t is:

$$\mathfrak{L}_{\text{TH}}(m, t) = \begin{cases} 1 \text{ if } t > \theta_m \text{ and } m \text{ is monotone increasing;} \\ 1 \text{ if } t < \theta_m \text{ and } m \text{ is monotone decreasing;} \\ 0 \text{ otherwise.} \end{cases}$$

The thresholds are treated as free parameters in the model, which are to be inferred from the data. We use Bayesian inference to encode prior expectations about the likely meanings of WEPs as weakly informative prior distributions over thresholds (Gelman, Carlin, Stern, & Rubin, 2014).

Prototype-based lexicon

The prototype-based approach holds that WEPs denote gradient and focalised ranges on the probability scale. To implement this approach, we make use of *fuzzy logic*, which argues that the truth value of a sentence can take any value in the [0, 1] interval (e.g., Zadeh, 1983, 1996). Specifically, we assume that each WEP is associated with two parameters: a *prototype* and a *distance measure*.

The prototype is the state of affairs in which a WEP is maximally true. The distance measure modulates the effect of distance from the prototype on the truth value of the WEP. This distance measure captures the intuition that WEPs vary in their strictness, e.g., intuitively, 'impossible' requires that the probability be very close to 0%, whereas 'possible' is felicitous in a much wider range of situations.

Thus, we define a prototype-based lexicon $\mathfrak{L}_{\text{PT}}$. This lexicon associates each message m with a prototype p_m and a distance measure d_m, so that the truth value of m in state t is:

$$\mathfrak{L}_{\text{PT}}(m, t) = exp\left(-\left(\frac{t - p_m}{d_m}\right)^2\right)$$

Similarly to the threshold-based lexicon, prototypes and distance measures are treated as free parameters in the model, to be inferred from the data.

Speaker models

Given a lexicon, we may define two types of speakers in order to connect the hypothesised semantics to the data from the production experiment: a *literal* speaker and a *pragmatic* speaker. The literal speaker solely aims at being truthful, i.e., she prefers to produce true messages over false ones ($\mathfrak{L}_{\text{TH}}$), or messages with a higher truth value over messages with a lower truth value ($\mathfrak{L}_{\text{PT}}$).

We further assume that the available messages vary in their *salience*. Some WEPs come to mind more easily than others, as evidenced by their fluctuating production frequencies. To model effects of differential salience, we pair each message m with a salience value $P_{\text{Sal}}(m)$, which is treated as a free variable.

Thus, we may define a literal speaker S_{lit} as follows:

$$P_{S_{\text{lit}}}(m \mid t, \mathfrak{L}) \propto P_{\text{Sal}}(m)\ \mathfrak{L}(m, t)$$

This definition states that the probability that the literal speaker produces a message m in a state of affairs t is proportional to (i) the salience of m, and (ii) the truth value of m in t.

The literal speaker only cares about truthfulness. However, one of the central tenets of modern pragmatics is that speakers tend to behave *rationally*, i.e., optimise the probability that their audience arrives at the correct interpretation

(Grice, 1975). Given people's cognitive limitations, we may plausibly expect that this form of rationality is bounded, so that speakers prefer to select the optimal message but occasionally deviate from this optimum.

Accordingly, we may define a pragmatic speaker S_{prag}. The pragmatic speaker is truthful but also seeks to optimise the chance of coordination with a literal listener L_{lit}. The literal listener, in turn, naively infers a state of affairs with a probability that is proportional to the truth value of the message in that state:

$$P_{S_{\text{prag}}}(m \mid t, \mathfrak{L}) \propto P_{\text{Sal}}(m)\ P_{L_{\text{lit}}}(t \mid m, \mathfrak{L})^{\lambda}, \text{where}$$
$$P_{L_{\text{lit}}}(t \mid m, \mathfrak{L}) \propto \mathfrak{L}(t, m)$$

The lambda parameter λ modulates the probability with which the speaker chooses the pragmatically optimal message, i.e., the message that is the most likely to receive the intended interpretation on the part of the listener (cf. Zaslavsky, Hu, & Levy, 2020). We later investigate whether the inferred value for this parameter varies between participants with a high and low AQ.

The literal and pragmatic speaker models specify the probability of a message given a state. Our linking hypothesis is that this probability reflects the probability that a speaker would produce that message in that state, i.e., our hypothesis is that it approximates the corresponding production probability in Exp. 1.

The current speaker models assume perfect knowledge of the actual state of affairs. Though there may be circumstances in which this assumption is plausible, the use of WEPs is generally associated with uncertainty about the actual probability (e.g., Teigen & Brun, 2003). To model such uncertainty, we enrich the speaker models with a module representing the approximate perception of numerosity. While there may be other factors that cause uncertainty, the inaccurate perception of the number of red marbles is presumably the most prominent one in the context of our experiment.

Number perception

Speakers in the production experiment had to estimate the actual probability of drawing a red marble, i.e., the number of red marbles. The cognitive system used to estimate large numerosities is called the *Approximate Number System* (ANS) (Dehaene, 1997; Feigenson, Dehaene, & Spelke, 2004). It is well known that the estimates of the ANS are prone to error. In particular, the accuracy of the ANS decreases as the number to be estimated increases.

To model the accuracy of participants' estimates, we define the confusion probability $P_{\text{Cf}}(t' \mid t)$ of perceiving the actual state of affairs t as t'. Since the visual displays in the production experiment were upper-bounded, $P_{\text{Cf}}(t' \mid t)$ is defined as the product of the probability $P_{\text{ANS}}(t' \mid t)$ of maintaining an approximate representation of the number t as t' and the inverse probability $P_{\text{ANS}}(100 - t \mid 100 - t')$. These probabilities, in turn, are specified as follows:

$$P_{\text{Cf}}(t' \mid t) \propto P_{\text{ANS}}(t' \mid t)\ P_{\text{ANS}}(100 - t' \mid 100 - t)$$

$$P_{\text{ANS}}(t' \mid t) = \int_{t'-0.5}^{t'+0.5} \text{Gaussian}(x, \mu = t, \sigma = w\ t)\text{d}x$$

The parameter w stands for *Weber's fraction*, which represents the accuracy of participants' estimates. To parametrise w, we carried out an experiment (Exp. 2) in which we presented 50 participants with the same types of displays used in the production experiment (Fig. 2).[3] Participants had to estimate the percentage of red marbles using a continuous slider. Each participant saw 25 vases with random proportions of red marbles. Based on the results of this experiment, we determined that the maximum likelihood estimate of the Weber fraction was $\hat{w} = 0.35$. We use this value in all production models.

We added the numerosity estimation module to our speaker models. If $P_S(m \mid t, \mathfrak{L})$ is a speaker production rule, either literal or pragmatic, the production

[3] See the Appendix for a more detailed description of the numerosity estimation experiment.

probabilities under approximate perception of the actual state are:

$$P_S^{\mathrm{Cf}}(m \mid t, \mathfrak{L}) \propto \sum_{t' \in T} P_{\mathrm{Cf}}(t' \mid t)\ P_S(m \mid t', \mathfrak{L})$$

Model comparison

Taken together, we may distinguish four speaker models, varying the lexicon between threshold-based and prototype-based, and the speaker type between literal and pragmatic.[4] All four models were implemented in Stan (Stan Development Team, 2018) to obtain samples from the posterior distribution over free parameter values conditioned on the data from the production experiment. For the purpose of model comparison, we split our dataset into a training set (the first 195 participants) and a test set (the remaining 60 participants). The four models were trained on the larger training set and were evaluated based on how well they explained the training set.

Fig. 4 shows the data and posterior predictive distribution of the models. This figure suggests that the literal threshold-based model offers a relatively poor approximation of the production data, while the other three models fare much better.

For proper statistical model comparison, we look at how well each model is able to predict the test dataset by calculating the expected log pointwise predictive density using the 'elpd()' function from the R package 'loo' (Vehtari et al., 2022). For this analysis, we used the unbinned data and predictions. The model fits were statistically compared using the 'loo_compare()' function from the same package. Table 1 shows the outcome of this comparison.

[4]Note, in passing, that if we construe prototype-based meanings as reflections of patterns in language use, it seems redundant to assume that they enter into a further pragmatic reasoning

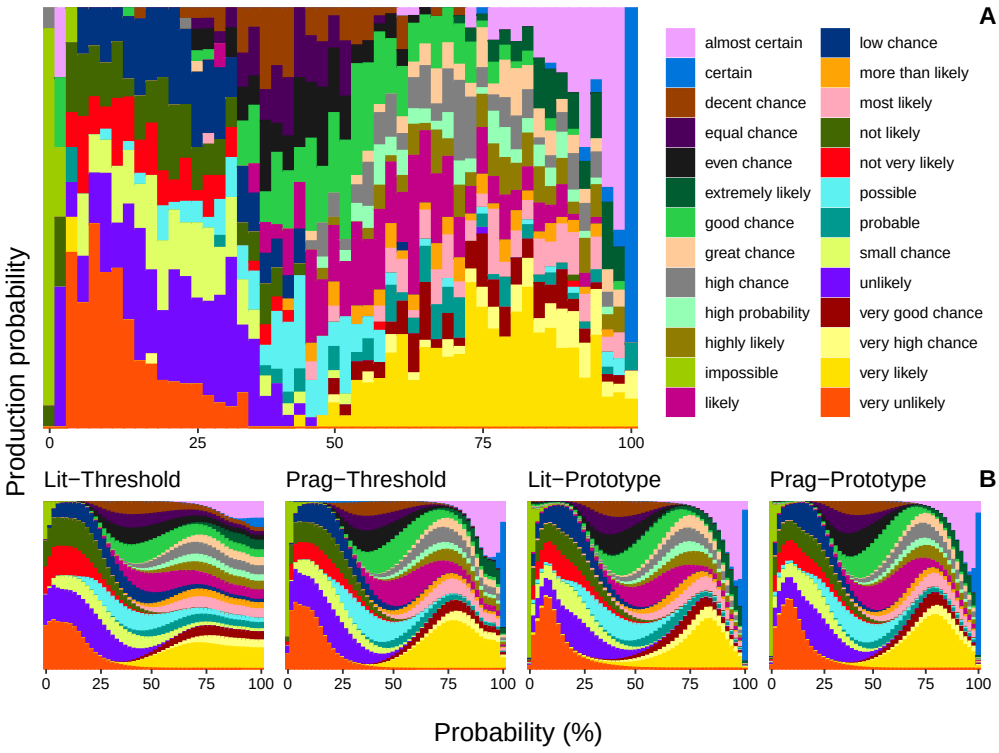

Figure 4: **A**: Production probabilities of the WEPs in our sample (Exp. 1). **B**: Predicted production probabilities for each of the four speaker models. Note that, for this figure, production probabilities are binned into bins consisting of two adjacent probabilities, except for probabilities 0 and 100.

	elpd_diff	se_diff
prag-prototype	0.0	0.0
prag-threshold	-6.5	8.0
lit-prototype	-13.9	5.1
lit-threshold	-63.4	15.4

Table 1: Differences in expected log pointwise predictive density relative to the optimal pragmatic prototype-based model (elpd_diff) and corresponding standard error of the difference (se_diff). The expected log pointwise predictive density is a measure of overall model fit, so the difference indicates how much worse the model predictions are compared to the pragmatic prototype-based model.

The table indicates that the pragmatic prototype-based model was the optimal one, but was not significantly better than the pragmatic threshold-based model, since the difference is smaller than corresponding standard error. By contrast, the pragmatic prototype-based model was superior to the other two models, since the difference in both cases is greater than twice the standard error. The comparable fit of the two pragmatic models shows that patterns of gradience and focality in the use of WEPs can be explained equally well within a threshold-based approach as within a prototype-based approach that directly encodes these patterns into the semantics of WEPs.

Before discussing these results, we turn to investigate the effects of AQ on the lambda parameter that modulates the probability with which the speaker chooses the pragmatically optimal message.

process. Nevertheless, for completeness, we also consider the possibility of a pragmatic speaker using prototype-based meanings.

Autism spectrum quotient

After the production experiment, we asked all participants to fill out the Autism Spectrum Quotient test, which is a 50-question multiple choice questionnaire in which participants have to indicate if they agree or disagree with certain statements that pertain to traits that are often associated with ASD.

To investigate the effect of AQ on speaker behaviour, we divided participants based on whether their AQ was above or below the median AQ across all participants in our sample (i.e., 22), with participants at the median assigned to the high-AQ group. The average AQ of the high-AQ group was 26 (range: 22–36); of the low-AQ group 14 (range: 3–21). Due to computational limitations, we could not incorporate AQ as a continuous measure; hence, this analysis is inevitably coarse-grained.

To put these AQ values in perspective, Baron-Cohen et al. (2001) suggest that an AQ of 32 or higher is a reliable indicator of the presence of Autism Spectrum Disorder (ASD), since approximately 80% of their autistic participants had an AQ of at least 32, compared to only 2% of their neurotypical participants. Our high-AQ participants were mostly below this threshold, i.e., even though they exhibited autistic traits, they were generally not at risk of having ASD.

For this analysis, we focus on the pragmatic threshold-based model. Using STAN, we obtained samples from the posterior distribution over free parameter values conditioned on the data from the production experiment. We fit the model using the combined training and test datasets. Crucially, we fit different lambda parameters for the datasets from high-AQ and low-AQ participants.

By fitting the model on the entire dataset, rather than on the datasets from high-AQ and low-AQ participants separately, we ensure that all parameters except for the lambda parameter remain constant across both groups of participants, so that

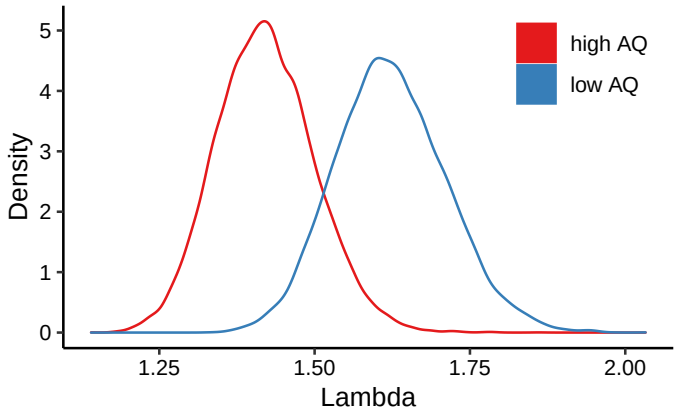

Figure 5: Density plot of the posterior estimates of the lambda parameter for high-AQ and low-AQ participants.

differences in the lambda parameter cannot be interpreted as statistical "spandrels" compensating for differences in other parameters. Reassuringly, the same pattern of results emerges if the model is in fact fit on both datasets separately.

Fig. 5 shows the posterior estimates of the lambda parameter for high-AQ and low-AQ participants. The mean estimated lambda parameter for high-AQ participants was 1.42; for low-AQ participants 1.62. A t-test indicated that this difference was significant ($t(31519) = 215$, $p < .001$). This analysis suggests that participants with a high AQ were less likely to select the pragmatically optimal message than participants with a low AQ.

General discussion

People associate WEPs with gradient and focalised ranges on the probability scale. It has sometimes been concluded that these patterns of use must be reflected in the underlying semantics of WEPs (e.g., Bocklisch et al., 2012; Jaffe-Katz et al., 1989). Here, we have shown that this conclusion is unwarranted: data from a novel production experiment could be explained equally well on the basis of a truth-

conditional approach that associates WEPs with crisp thresholds on the probability scale as on the basis of a prototype-based approach that directly encodes gradience and focality into the semantics of WEPs. Importantly, this equivalence only holds if the threshold-based approach is embedded in a probabilistic model that encodes perceptual limitations and goal-directed speech.

On a more abstract level, our results lend support to a modular approach that distinguishes between the conventional meaning of an expression and what a speaker who uses that expression conveys. This bifurcation between meaning and use has a number of important theoretical advantages (e.g., explaining entailment patterns, cf. Barwise & Cooper, 1981). In this paper, we provide further support by showing precisely how the lean meanings postulated by the threshold-based approach may lead to rich and complex patterns in language use.

One of the central insights of modern pragmatics is that speaker behaviour can be viewed, to a large extent, as rational, i.e., goal-oriented action. Recent probabilistic models provide a means to precisely quantify to what extent speakers behave rationally. Interestingly, Autism Spectrum Disorder (ASD) is said to be characterised, in part, by a pragmatic deficit. Hence, we intuited that this deficit might be reflected in the model parameters, specifically in a parameter that modulates the degree of rationality. We indeed find that participants with more autistic traits—as measured using the Autism Spectrum Quotient test—were estimated to have a significantly lower rationality parameter than participants with fewer autistic traits.

This observation provides an interesting counterpoint to earlier findings showing that participants with and without ASD are equally likely to derive *scalar inferences*, such as the inference from ‘some’ to ‘not all’ (e.g., Chevallier, Wilson, Happé, & Noveck, 2010; Pijnacker, Hagoort, van Buitelaar, Teunisse, & Geurts, 2009; Su & Su, 2015). The derivation of scalar inferences is also assumed to

be reliant on (the hearer's assumption of) rational speaker behaviour (e.g., Geurts, 2010; Horn, 1972). How can this discrepancy be explained, i.e., why do people with and without ASD derive scalar inferences at equivalent rates but do people with more autistic traits behave less rationally in our production experiment? One possible explanation is that the pragmatic effects of autistic traits are too subtle to be brought out using coarse-grained tasks such as asking whether sentences like 'Some dogs are mammals' are true or false, but that these effects surface in more naturalistic contexts as exemplified by our production experiment (cf. van Tiel & Kissine, 2018).

At the same time, it should be noted that the connection between AQ and ASD is not uncontentious. First, a number of authors have argued that the AQ test is not an adequate predictor of the presence or absence of ASD (e.g., Ashwood et al., 2016; Lundqvist & Lindner, 2017). In particular, these studies show that the AQ thresholds used for identifying people at risk of having ASD substantially underestimate the actual prevalence of ASD. More problematically, it has recently been argued that there are important theoretical and practical problems associated with the construal of autism as a spectrum, i.e., as a collection of traits that are to a lesser degree also shared by the non-autistic population (Mottron & Bzdok, 2020; Sasson & Bottema-Beutel, 2022). Given these concerns, the current findings call for confirmation using gold-standard instruments such as the Autism Diagnostic Observation Schedule (Lord et al., 2000).

Our production experiment elicited various types of WEPs, including adjectival (e.g., 'likely') and nominal (e.g., 'a good chance') ones. The pragmatic model assumes that these WEPs compete with each other to the same degree. At the same time, we observed that about 40% of the participants consistently produced either adjectival or nominal WEPs. Consequently, adjectival WEPs were more likely to co-occur with other adjectival WEPs than with nominal ones, and vice

versa, suggesting that expressions from the same part of speech compete with each other more strongly than with expressions from different parts of speech. An interesting direction for future research is to encode such differential levels of "alternativeness" into the model.

The communication of probability is of great importance in high-risk areas such as healthcare (e.g., Lipkus, 2007). Here, we have successfully implemented a computational model that explains the use of probability expressions while being firmly rooted in sound linguistic theory. We hope to have thereby contributed to a better understanding of the use and misuse of these expressions.

References

Alstott, J., & Jasbi, M. (2020). Lexicalization of quantificational forces in adverbial and determiner domains. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society* (pp. 2001–2006).

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Author.

Ashwood, K. L., Gillan, N., Horder, J., Hayward, H., Woodhouse, E., McEwen, F. S., … Murphy, D. G. (2016). Predicting the diagnosis of autism in adults using the autism-spectrum quotient (AQ) questionnaire. *Psychological Medicine*, *46*, 2595—2604.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. MIT Press.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism spectrum quotient (AQ): Evidence from Asperger syndrome / high-functioning autism, males and females, scientists and mathematicians.

Journal of Autism and Developmental Disorders, *31*, 5–17.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, *4*, 159–219.

Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. *Journal of Forecasting*, *1*, 257–269.

Bocklisch, F., Bocklisch, S. F., & Krems, J. F. (2012). Sometimes, often, and always: Exploring the vague meanings of frequency expressions. *Behavior Research Methods*, *44*, 144–157.

Budescu, D. V., & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. *Psychology of Learning and Motivation*, *32*, 275–318.

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 281–294.

Chevallier, C., Kohls, G., Troiani, V., Brodkin, E. S., & Schultz, R. T. (2012). The social motivation theory of autism. *Trends in Cognitive Sciences*, *16*, 231–239.

Chevallier, C., Wilson, D., Happé, F., & Noveck, I. (2010). Scalar inferences in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *40*, 1104–1117.

Clark, H. H. (1990). Comment on Mosteller and Youtz "Quantifying probabilistic expressions". *Statistical Science*, *5*, 12–16.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. Oxford University Press.

Elsaesser, C., & Henrion, M. (1990). How much more probable is "much more probable"? Verbal expressions for probability updates. In M. Henrion, R. D. Shachter, L. M. Kanal, & J. F. Lemmer (Eds.), *Uncertainty in artificial intelligence 5* (pp. 319–328). Elsevier.

Erev, I., & Cohen, B. L. (1990). Verbal versus numerical probabilities: Efficiency, biases and the preference paradox. *Organizational Behavior and Human Decision Processes*, *45*, 1–18.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*, 307–314.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998.

Franke, M. (2009). *Signal to act: Game theory in pragmatics* (Unpublished doctoral dissertation). University of Amsterdam.

Friedman, J. A., & Zeckhauser, R. (2015). Handling and mishandling estimative probability: Likelihood, confidence, and the search for Bin Laden. *Intelligence and National Security*, *30*, 77–99.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Chapman.

Geurts, B. (2010). *Quantity implicatures.* Cambridge University Press.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Science*, *20*, 818–829.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics (vol. 3). Speech acts* (pp. 41–58). Academic Press.

Herbstritt, M., & Franke, M. (2019). Complex probability expressions & high-

order uncertainty: Compositional semantics, probabilistic pragmatics & experimental data. *Cognition*, *186*, 50–71.

Horn, L. R. (1972). *On the semantic properties of logical operators in English* (Unpublished doctoral dissertation). University of California, Los Angeles.

Jaffe-Katz, A., Budescu, D. V., & Wallsten, T. S. (1989). Timed magnitude comparisons of numerical and nonnumerical expressions of uncertainty. *Memory & Cognition*, *17*, 249–264.

Juanchich, M., & Sirota, M. (2019). Do people really prefer verbal probabilities? *Psychological Research*, *84*, 2325–2338.

Karelitz, T. M., & Budescu, D. V. (2004). You say "probable" and I say "likely": Improving interpersonal communication with verbal probability expressions. *Journal of Experimental Psychology: Applied*, *10*, 25–41.

Kent, S. (1964). Words of estimative probability. *Studies in Intelligence*, *8*, 49–65.

Kissine, M. (2012). Pragmatics, cognitive flexibility and autism spectrum disorders. *Mind & Language*, *27*, 1–28.

Kissine, M. (2021). Autism, constructionism and nativism. *Language*, *97*, 139–169.

Kratzer, A. (1991). Modality. In A. von Stechow & D. Wunderlich (Eds.), *Semantik: Ein internationales Handbuch der zeitgenössischen Forschung* (pp. 639–650). Walter de Gruyter.

Lassiter, D. (2019). *Graded modality: Qualitative and quantitative perspectives*. Oxford University Press.

Lichtenstein, S., & Newman, J. R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Science*, *9*, 563–564.

Lipkus, I. M. (2007). Numeric, verbal, and visual formats of conveying health risks: Suggested best practices and future recommendations. *Medical Decision Making*, *27*, 696–713.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., … Rutter, M. (2000). The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, *30*, 205–223.

Lundqvist, L.-O., & Lindner, H. (2017). Is the autism-spectrum quotient a valid measure of traits associated with the autism spectrum? A Rasch validation in adults with and without autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *47*, 2080–2091.

Moss, S. (2015). On the semantics and pragmatics of epistemic vocabulary. *Semantics and Pragmatics*, *8*, 1–81.

Mosteller, F., & Youtz, C. (1990). Quantifying probability expressions. *Statistical Science*, *5*, 2–34.

Mottron, L., & Bzdok, D. (2020). Autism spectrum heterogeneity: Fact or artifact. *Molecular Psychiatry*, *25*, 3178–3185.

Partee, B. (1999). Semantics. In R. A. Wilson & F. C. Keil (Eds.), *The MIT encyclopedia of the cognitive sciences* (pp. 739–742). MIT Press.

Partee, B. (2001). Montague grammar. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (pp. 9995–9999). Pergamon.

Pijnacker, J., Hagoort, P., van Buitelaar, J., Teunisse, J.-P., & Geurts, B. (2009).

Pragmatic inferences in high-functioning adults with autism and Asperger syndrome. *Journal of Autism and Developmental Disorders*, *39*, 607–618.

Reagan, R. T., Mosteller, F., & Youtz, C. (1989). Quantitative meanings of verbal probability expressions. *Journal of Applied Psychology*, *74*, 433–442.

Renooij, S., & Witteman, C. (1999). Talking probabilities: Communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning*, *22*, 169–194.

Sasson, N. J., & Bottema-Beutel, K. (2022). Studies of autistic traits in the general population are not studies of autism. *Autism*, *26*, 1007–1008.

Schuster, S., & Degen, J. (2020). I know what you're probably going to say: Listener adaptation to variable use of uncertainty expressions. *Cognition*, *203*, 104285.

Shinagare, A. B., Lacson, R., Boland, G. W., Wang, A., Silverman, S. G., Mayo-Smith, W. W., & Khorasani, R. (2019). Radiologist preferences, agreement, and variability in phrases used to convey diagnostic certainty in radiology reports. *Clinical Practice Management*, *16*, 458–464.

Stan Development Team. (2018). *The Stan core library.* (Version 2.18.0)

Su, Y., & Su, L.-Y. (2015). Interpretation of logical words in Mandarin-speaking children with autism spectrum disorders: Uncovering knowledge of semantics and pragmatics. *Journal of Autism and Developmental Disorders*, *45*, 1938–1950.

Swanson, E. P. (2006). *Interactions with context* (Unpublished doctoral dissertation). Massachussetts Institute of Technology.

Teigen, K. H., & Brun, W. (2003). Verbal expressions of uncertainty and probability.

In D. Hardman & L. Macchi (Eds.), *Thinking: Psychological perspectives on reasoning, judgment and decision making* (pp. 125–145). Wiley.

van Tiel, B., Franke, M., & Sauerland, U. (2021). Probabilistic pragmatics explains gradience and focality in natural language quantification. *Proceedings of the National Academy of Sciences*, *118*, e2005453118.

van Tiel, B., & Kissine, M. (2018). Quantity-based reasoning in the broader autism phenotype: A web-based study. *Applied Psycholinguistics*, *39*, 1119–1154.

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2022). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models.* Retrieved from https://mc-stan.org/loo/ (R package version 2.5.1)

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General*, *115*, 348–365.

Willems, S., Albers, C., & Smeets, I. (2020). Variability in the interpretation of probability phrases used in Dutch news articles — a risk for miscommunication. *Journal of Science Communication*, *19*, 1–26.

Yalcin, S. (2007). Epistemic modals. *Mind*, *116*, 983–1026.

Zadeh, L. A. (1983). A computational approach to fuzzy quantifiers in natural languages. *Computers & Mathematics with Applications*, *9*, 149–184.

Zadeh, L. A. (1996). *Fuzzy sets, fuzzy logic, and fuzzy systems: Selected papers* (G. J. Klir & B. Yuan, Eds.). World Scientific.

Zaslavsky, N., Hu, J., & Levy, R. (2020). *A rate-distortion view of human pragmatic reasoning.* arXiv. Retrieved from https://arxiv.org/abs/2005.06641

Zimmer, A. C. (1983). Verbal versus numerical processing of subjective probabilities. In R. W. Scholz (Ed.), *Decision making under uncertainty* (pp. 159–182). North Holland.