Bespoke probabilistic modelling reveals subtle effects of autism on pragmatic optimisation in the expression of quantification

Bob van Tiel[1], Michael Franke[2], Uli Sauerland[3], and Philippine Geelhand[4]

[1]Radboud University Nijmegen, The Netherlands

[2]University of Tübingen, Germany

[3]Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin, Germany

[4]Université Libre de Bruxelles, Belgium

(to appear in *Computational Brain & Behavior*)

Statements and declarations

Funding: The work of MF and US has benefitted from the project LMBayes (grant: Leibniz Collaborative Excellence 22-20, PI Anton Benz). The work of PG was supported by a F.R.S-FNRS Postdoctoral Fellowship (grant number: 40000037).

Conflict of interest: The authors declare no conflict of interest.

Availability of data and code: The anonymised data and analysis files associated with this article can be accessed via the following link:

https://osf.io/gfxde/overview?view_only=6acf16bd2c144ecf84e3160477fc66f5

Authors' contributions: All authors formulated the project, and designed the experiments. PG conducted the experiments. MF developed the probabilistic model. BvT and MF implemented and tested the model. All authors wrote the paper.

Ethics approval and consent: All procedures received written approval from the ethical committee of the Université Libre de Bruxelles in accordance with the 1964 Declaration of Helsinki and its later amendments.

Consent: Informed consent was obtained from all participants included in the study.

Address for correspondence: Correspondence concerning this article should be addressed to Bob van Tiel, Radboud University, Faculty of Philosophy, Theology, and Religious Studies, Postbus 9102, 6500 HC Nijmegen, The Netherlands. Email: bobvantiel@gmail.com

Abstract

Autism is often associated with difficulties with pragmatic language use. However, several studies on the interpretation of quantifying expressions (e.g., 'some') failed to observe differences between autistic and neurotypical participants. In this study, we present data from a more naturalistic free-production study comparing how French-speaking autistic and neurotypical adults use quantifying expressions. To capture the complexity of this approach, we employ a bespoke probabilistic model that estimates a theoretically motivated pragmatic optimality parameter, which we take as a measure of pragmatic ability. Our results show that, as a group, autistic participants are less likely to produce optimally informative quantifying expressions than their neurotypical peers. Nonetheless, the difference between the two populations is relatively small, pointing to a difference in degree rather than kind. These findings contribute to ongoing debates about pragmatic ability in autism, and highlight the value of bespoke probabilistic modelling for testing hypotheses about subtle differences in latent traits and their complex empirical signatures.

Keywords: computational model; autism; pragmatics; quantification

Introduction

Autism spectrum disorder is a lifelong neurobiological condition that is characterised, on the one hand, by restrictive and repetitive patterns of behaviour, interests, and activities, and on the other hand, by pervasive difficulties with social communication and interaction (American Psychiatric Association, 2013; World Health Organisation, 2022). As the term 'spectrum' implies, autism is highly heterogeneous in many respects, including developmental history, intelligence, comorbidity, and symptom severity (e.g., Masi, DeMayo, Glozier, & Guastella, 2017; Mottron & Bzdok, 2020).[1]

This heterogeneity also extends to language abilities, which range from being completely nonverbal to having unimpaired structural language skills (Kim, Paul, Tager-Flusberg, & Lord, 2014; Volden, Coolican, Garon, White, & Bryson, 2009). However, even autistic people with structural language skills in the normal range experience difficulties with *pragmatics*, i.e., aspects of language use in context, although again the scope and severity of these difficulties varies greatly both across individuals and across pragmatic phenomena (e.g., Geurts, Kissine, & van Tiel, 2019; Jary, Martín-González, Vicente, &

[1] A note on terminology: we recognise that the autism community encompasses diverse viewpoints and identities, and we respect the right of all individuals with lived experience of autism to choose language that reflects how they identify. Based on preference data from French-speaking autistic adults (Geelhand et al., 2023), we use identity-first language (e.g., 'autistic person') throughout this paper.

Castroviejo, 2025; Tager-Flusberg, Paul, & Lord, 2005; Vicente & Falkum, 2023). Often, the observed pragmatic difficulties are connected to more general difficulties with *mind-reading*, i.e., the ability to accurately determine what someone else thinks, intends, or desires (e.g., Baron-Cohen, 2013; Leslie & Thaiss, 1992), though, again, autistic people vary greatly in the extent to which they experience such difficulties (e.g. Begeer, Rieffe, Meerum Terwogt, & Stockmann, 2003; Gernsbacher & Yergeau, 2020; Marrocchini, 2023).

A particularly interesting case of pragmatic language use concerns the appropriate choice of *quantifying expression* (e.g., 'few', 'some', 'most', 'all'). In formal semantics, quantifying expressions are often analysed as expressing relations between sets (e.g, Barwise & Cooper, 1981; Montague, 1973). For example, 'Some S are P' is taken to mean that the sets S and P denoted by the subject and the predicate share at least one element, i.e., $|S \cap P| \geq 1$. Other examples of such logical definitions are given in (1).

(1) a. 'Few S are P' means $|S \cap P| \leq \delta|S|$

b. 'Most S are P' means $|S \cap P| > \frac{1}{2}|S|$

c. 'All S are P' means $|S \cap P| = |S|$

Here, δ is a context-dependent parameter, which reflects the observation that the interpretation of expressions such as 'few' relies on the content and context of the utterance (e.g.,

Lorson, Macuch-Silva, Hart, & Winter, 2024; Moxey & Sanford, 1993; Schöller & Franke, 2017).

Although these logical definitions adequately capture the *meanings* of quantifying expressions, they do not straightforwardly connect to the way people *use* them (e.g., Ramotowska, 2022; Schlotterbeck, Ramatowska, van Maanen, & Szymanik, 2020). To illustrate, consider a situation in which 100 out of 100 students passed the exam. Based on their logical meanings, it would be truthful to say that *some*, *most*, or *all* of the students passed. However, most people would naturally choose to say that *all* of the students passed the exam. A standard explanation for this behavioral regularity is that speakers aim not merely to be truthful, but also to be *informative* (Grice, 1975). Because the logical meaning of 'all' is stronger than that of 'some' or 'most'—in the sense that it rules out a larger set of possible states—an informative speaker is predicted to prefer 'all' whenever all three expressions are true.

While accounts of pragmatic reasoning were long confined to the domain of verbal theories, more recently there has been a growing interest in formal and computational modelling of pragmatic reasoning (e.g., Degen, 2023), in conjunction with experimental data (e.g., Noveck, 2018). In this paper, we explore the potential of such data-driven probabilistic modelling to add nuance to the debate about the pragmatic ability of autistic people. Concretely, we build on a previous model of the production of quantifying expres-

sions (van Tiel, Franke, & Sauerland, 2021), and previous exploratory results showing a relation between the autism spectrum quotient of neurotypical adults and the degree to which more informative (i.e., pragmatically optimal) expressions are chosen (van Tiel, Sauerland, & Franke, 2022). Here, we present novel data specifically comparing autistic and neurotypical participants. We show how bespoke probabilistic modeling enables a direct comparison of the degree of pragmatic optimisation across the two participant groups.

The paper is structured as follows. The next section motivates our approach more thoroughly against the background of related prior work, while paying special attention to the potential benefits of computational modeling in this domain. We then detail the computational model, describe our experiments and empirical data, and finally report on the results of the model-based data analysis.

Autism and pragmatic reasoning about informative quantifying expressions

The contextually adequate use of quantifying expressions often requires reasoning about how others will interpret what was said. On traditional accounts of autism, one might therefore expect autistic individuals to encounter difficulties with pragmatic reasoning about informativeness. Yet, numerous studies have failed to support this prediction (e.g., Chevallier, Wilson, Happé, & Noveck, 2010; Hochstein, Bale, & Barner, 2017; Pijnacker,

Hagoort, van Buitelaar, Teunisse, & Geurts, 2009; Schaeken, Van Haeren, & Bambini, 2018; Su & Su, 2015; van Tiel & Kissine, 2018, but see Mazzaggio & Surian, 2018; Pastor-Cerezuela, Tordera Yllescas, González-Sala, Montagut-Asunción, & Fernández-Andrés, 2018 for evidence to the contrary).

To illustrate, Pijnacker and colleagues asked autistic and neurotypical participants to judge whether statements such as (2) were true or false.

(2) Some dogs are mammals.

Logically, this statement is true: there exist dogs that are mammals. Pragmatically, however, it suggests that not all dogs are mammals. This pragmatic inference is justified by considerations of informativeness: given that all dogs are mammals, the optimally informative utterance would have been (3) rather than (2).

(3) All dogs are mammals.

By opting for the weaker statement, the speaker seems to convey the false belief that not all dogs are mammals. Accordingly, 'false' responses indicate that participants drew the pragmatic inference. Pijnacker and colleagues found no significant difference between autistic and neurotypical participants in the rate at which such statements were rejected, and concluded that autistic people show no problems with pragmatic reasoning about

informativeness; a finding that was replicated in several subsequent studies (e.g., Chevallier et al., 2010; Su & Su, 2015).

These results are informative, but it is debatable whether they fully support the conclusion. First, truth-value judgement tasks rely heavily on *metalinguistic reasoning*: participants are explicitly asked to reflect on the appropriateness of linguistic expressions. Such reflective judgements may mask difficulties that would emerge in more spontaneous language use (see Khorsheed & van Tiel, 2024, for a similar argument focusing on bilingualism). Second, these experiments typically involve simple binary contrasts, such as 'some' versus 'all'. In natural language, however, quantifying expressions compete in a far richer space, with alternatives such as 'most', 'many', 'half', or 'very few' also available (e.g., Horn, 1972, 1989; van Tiel et al., 2021). Autistic individuals may have little trouble with simple pairs but face greater challenges when weighing the communicative utility of expressions in a more complex lexical space.

Van Tiel et al. (2022) provide preliminary evidence for this possibility. Their study examined probability expressions, such as 'likely' and 'certain'. In the semantics literature, such expressions are typically analysed as involving quantification over probabilities (e.g., Lassiter, 2019). In a production task, neurotypical participants were shown urns containing 100 marbles with varying proportions of red marbles and asked to describe the probability of drawing a red marble by completing the sentence frame 'If you randomly

draw a marble from this urn, — that it is red'. Crucially, this task required participants to select an expression from the full space of probability expressions, rather than from a simple two-item contrast.

Participants also completed the autism spectrum quotient (AQ) questionnaire, a 50-item instrument that produces a score reflecting the extent to which respondents indicate that they have traits associated with ASD (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001). Based on these scores, participants were grouped into low-AQ and high-AQ categories. Van Tiel and colleagues analysed the production data with a computational model of quantifying expression choice introduced in earlier work (van Tiel et al., 2021). The model estimates the probability that participants select informative descriptions. We will describe this model in more detail later, but what is relevant here is that the model provides an estimate of the probability with which participants produce optimally informative descriptions, i.e., a measure of the extent to which they behave pragmatically. Applying this model to their production data, van Tiel and colleagues found that the low-AQ group was significantly more likely to select informative descriptions than the high-AQ group.

While these findings are suggestive, they must be interpreted cautiously. The study did not compare autistic and neurotypical individuals directly, but instead contrasted participants with low versus high AQ scores. This approach has two limitations: (i) the

AQ's reliability as a measure of autistic traits has been questioned (e.g., Ashwood et al., 2016; Lundqvist & Lindner, 2017), and (ii) results from comparisons across AQ scores may not generalise to comparisons between autistic and neurotypical groups (e.g., Mottron & Bzdok, 2020; Sasson & Bottema-Beutel, 2022).

Together, these considerations highlight several desiderata for further research. First, to avoid potential masking effects of metalinguistic judgements on simple contrasts, it is preferable to use a more naturalistic free-production task. Second, direct comparisons between autistic and neurotypical participants are needed, rather than relying on AQ scores. Third, given that pragmatic differences in autism may be graded rather than categorical, analyses should be sensitive to degrees of pragmatic ability.

The latter is where bespoke computational modelling enables us to trace a measure of the latent degree of pragmatic optimisation. Concretely, the added benefit of computational modelling is that, when comparing an autistic to a neurotypical group based on data from a discrete choice task, we do not just want to test for arbitrary differences in the categorical choice distribution, but specifically for shifts in the response distribution that can be attributed to the relevant latent construct of pragmatic ability that we hypothesise—based on the previous literature—is responsible for these differences. It is here that data-driven computational modeling may be able to make a strong methodological contribution to an intricate empirical issue.

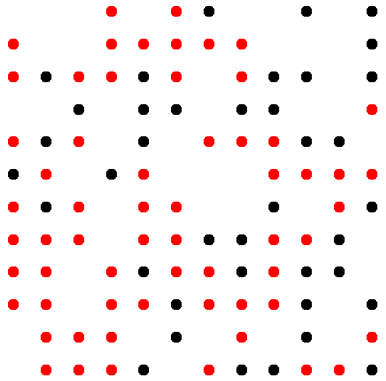

Figure 1: Example display used in our study.

With these aims in mind, we compare autistic and neurotypical participants' use of quantifying expressions in a free-production paradigm modelled on van Tiel et al. (2021) and van Tiel et al. (2022). Using their expression-choice model, we infer credible values of a latent pragmatic optimality parameter, thereby directly quantifying potential group differences in pragmatic ability.

Specifically, we report results from three new experiments. In Exp. 1, autistic and neurotypical participants described displays containing 100 red and black circles (see Fig. 1), by freely completing the sentence frame '— are red'.

Following van Tiel et al. (2021), we conducted two additional experiments to gather information relevant to the computational model. Exp. 2 examined the logical properties of quantifying expressions. Specifically, participants evaluated the logical validity of

arguments such as (4) and (5), deciding whether the truth of the premise necessarily guaranteed the truth of the conclusion.

(4) *Premise:* Some of the people ate salmon.

Conclusion: Some of the people ate fish.

(5) *Premise:* Some of the people ate fish.

Conclusion: Some of the people ate salmon.

The results of this experiment allowed us to determine whether particular quantifying expressions in our sample introduced a lower bound on the number of red circles (e.g., 'some' meaning '*at least* one'), or an upper bound (e.g., 'not all' meaning '*fewer than* all').

Exp. 3 focused on numerosity estimation. Participants were shown the same displays of red and black circles used in the production task (Fig. 1), and were asked to estimate the exact number of red circles. These data were used to estimate *Weber's fraction*, which is a standard measure of the accuracy of the approximate number system (ANS)—the cognitive mechanism responsible for estimating large numerosities (e.g., Dehaene, 1997). Since participants in the production task necessarily relied on the ANS to estimate numerosity, specifying Weber's fraction was essential for interpreting their performance.

Like Exp. 1, both Exps. 2 and 3 also tested both autistic and neurotypical participants. However, unlike in the production task, we did not anticipate group-level behavioural differences in these additional tasks.

Before turning to the details of the three experiments, we first provide an overview of the computational model that we applied to the production data.

A model for the production of quantifying expressions

Prolegomena

The model assumes a world with 101 possible states $T = \{t_0, ..., t_{100}\}$, corresponding to the 101 possible *intersection set sizes* $N = \{0, ...,100\}$. The intersection set size represents the number of red circles. The model also assumes 16 possible messages $M = \{m_{\text{all}}, ..., m_{\text{very few}}\}$. The selection of messages corresponds to the quantifying expressions that were the most frequently produced in a pretest that was similar to the main production experiment (see 'Exp. 1: Production' for details).

The model predicts which messages speakers produce in a given state on the basis of four ingredients: (i) a lexicon that links messages to their logical meanings, (ii) the salience of messages, (iii) numerosity perception, i.e., how participants perceive the intersection set size, and (iv) the speaker's degree of "pragmaticality". We now describe these ingredients in detail.

Lexicon

The lexicon associates each message m with a threshold θ_m on the intersection set size. This threshold can be either a lower bound or an upper bound, depending on the *monotonicity* of the quantifying expression. To illustrate, compare the following arguments:

(6) *Premise:* All of the people ate salmon.

Conclusion: All of the people ate fish.

(7) *Premise:* All of the people ate fish.

Conclusion: All of the people ate salmon.

The argument in the first example is valid: it is impossible for the premise to be true while the conclusion is false. In contrast, the second argument is invalid, as the premise could be true while the conclusion is false (e.g., if everyone ate tuna rather than salmon). Replacing 'all' with 'none' reverses the validity pattern.

Quantifying expressions that follow the inference pattern of 'all' are *monotone increasing*, whereas those that follow the pattern of 'none' are *monotone decreasing* (e.g., Barwise & Cooper, 1981). Monotone increasing quantifying expressions, such as 'all', introduce a lower bound on the intersection set size, while monotone decreasing quantifying expressions, such as 'none', introduce an upper bound. Following van Tiel et al. (2021), we

empirically determined the monotonicity profiles of the quantifying expressions used in our study (Exp. 2).

Formally, we define a lexicon $\mathfrak{L}$ that associates each message m with a threshold θ_m, so that the binary truth value of m in state t is:

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

Because people's intuitions often differ from logical definitions (e.g., Krifka, 2002; Newstead & Griggs, 1984; van Tiel, 2014), thresholds were treated as free parameters to be inferred from the data (see below under 'Model fitting').

Salience

Quantifying expressions differ in how readily they come to mind. For example, 'most' is typically more salient than 'more than half', even if they are semantically equivalent. To model this, each message is associated with a salience value $P_{\text{Sal}}(m)$, which is treated as a free parameter (see 'Model fitting').

Numerosity perception

Participants estimated the number of red circles in the displays. To model perceptual accuracy, we define the confusion probability $P_{\text{Conf}}(t' \mid t)$ for representing the true intersection set size t as t'. Given the bounded displays, this probability is defined as:

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

Here, $P_{\text{ANS}}(t' \mid t)$ is defined following standard assumptions about the approximate number system (e.g., Dehaene, 1997):

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

In this definition, w is Weber's fraction, reflecting participants' sensitivity to differences in the intersection set size, which is empirically measured in Exp. 3.

Speaker models and pragmaticality

We now integrate the preceding components into a model of how speakers select quantifying expressions. Our approach is grounded in a Gricean conception of cooperative communication, according to which speakers are assumed to produce utterances that serve the conversational goal. In the present task, we operationalise this goal as *belief alignment*: speakers aim to bring the listener's beliefs into alignment with their own representation of which state obtains (here, the intersection set size). Cooperativity does

not require that the speaker's representation be perfectly accurate; it requires only that speakers attempt to convey what they themselves take to be the case.

We implement this objective probabilistically. Communication is successful to the extent that the listener assigns high posterior probability to the state as the speaker represents it. Correspondingly, an utterance is more informative insofar as it increases this probability. "Pragmaticality" then consists in selecting utterances that maximise expected communicative success under this objective.

We formalise these assumptions within the Rational Speech Act (RSA) framework (e.g., Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013), which provides a probabilistic reconstruction of Gricean reasoning. In RSA, listeners update their beliefs via Bayesian inference, and speakers choose utterances according to a utility function defined over communicative success, as measured by the posterior probability that the listener recovers the speaker-intended state.

First, we define a *literal listener* $\mathrm{L}_{\mathrm{Lit}}$. This listener interprets a message strictly according to its lexical meaning, assigning equal probability to all states consistent with the message:

$$P_{\mathrm{L}_{\mathrm{Lit}}}(t \mid m, \mathfrak{L}) \propto \mathfrak{L}(t, m)$$

For example, if the literal listener hears ‘Some of the circles are red’, they consider any situation with at least one red circle equally probable, while rejecting any situation with no red circles. Although this literal listener is overly simplistic, it provides a foundation for modelling a *pragmatic speaker*.

The pragmatic speaker $\mathrm{S}_{\mathrm{Prag}}$ chooses messages to maximise the chance that the literal listener infers the correct intersection set size. Formally:

$$P_{\mathrm{S}_{\mathrm{Prag}}}(m \mid t, \mathfrak{L}) \propto P_{\mathrm{Sal}}(m) P_{\mathrm{L}_{\mathrm{Lit}}}(t \mid m, \mathfrak{L})^{\lambda}$$

To illustrate, consider a situation in which *all* of the circles are red. If the pragmatic speaker emits ‘all’, the probability that the literal listener correctly infers that all of the circles are red is 1, since this is the only state of the world in which the message is true. By contrast, if the pragmatic speaker emits ‘more than half’, the probability is only 1/49, since this message is true whenever the number of red circles lies between 51 and 100. Consequently, the pragmatic speaker is more likely to send ‘all’ than ‘more than half’.

In the definition of the pragmatic speaker, $P_{\mathrm{Sal}}(m)$ captures how readily a speaker comes up with a message, and λ regulates the speaker’s tendency to choose informative messages. A higher λ means the speaker is more likely to select messages that have a

higher probability that the listener infers the correct intersection set size—this is what we refer to as the speaker's "pragmaticality".

In the pragmatic speaker model, the speaker is assumed to have a perfectly accurate internal representation of the relevant state t. To model perceptual inaccuracy, we relax this assumption and allow for the possibility that the speaker's internal representation may differ from the actual state. Specifically, we assume that the true state t may be internally represented as a nearby state t' with probability $P_{\text{Conf}}(t' \mid t)$. Production behaviour is then obtained by marginalising over these possible internal representations:

$$P_{\text{S}_{\text{Prag}}^{\text{Conf}}}(m \mid t, \mathfrak{L}) \propto \sum_{t' \in T} P_{\text{Conf}}(t' \mid t)\, P_{\text{S}_{\text{Prag}}}(m \mid t', \mathfrak{L})$$

In this formulation, the intended interpretation is defined relative to the speaker's internal belief state. The speaker remains cooperative in the sense defined above, aiming to align the listener's beliefs with their own representation, but that representation may itself be inaccurate due to perceptual noise.

We restrict our analysis to cooperative communication and do not model deception. A deceptive speaker would require a different utility specification, in which communicative success is defined in terms of inducing a belief that the speaker takes to be false.

This extended pragmatic speaker model thus captures both the speaker's drive for informative communication and the limits imposed by perceptual uncertainty. We apply it to analyse the production of quantifying expressions in Exps. 1–3 and to compare pragmatic behaviour across autistic and neurotypical participant groups.

Exp. 1: Production

Participants

We recruited 67 native French-speaking participants. Thirty-four had a prior clinical diagnosis of autism (20 females; mean age 40.6, SD: 11.9, range 22–65), and 33 did not (21 females; mean age 29.0, SD: 10.8, range 18–60). We asked the second group whether they had any psychiatric or medical conditions. One individual reported a history of epilepsy but indicated that they had not experienced any episodes for several years. Because it remains unclear whether epilepsy constitutes neurodivergence (e.g., Boehm, 2021; Parker, 2023), and because this case involved a latent form of the condition, we refer to the second group as 'neurotypical' rather than 'non-autistic' (or 'allistic').

Autistic participants were significantly older than neurotypical participants ($t(65) = 4.2$, $p < .001$). Although there is mixed evidence about the relationship between age and pragmatic ability (see Bartczak, 2017), the observed age difference is relatively modest and unlikely to substantially affect our results.

Autistic participants were recruited from the Autism in Context: Theory and Experiment (ACTE) register of volunteers, and non-autistic participants via online announcements. Inclusion criteria were: (i) age $\geq$ 18 years, (ii) global IQ $>$ 70, (iii) normal or corrected-to-normal vision and hearing, and (iv) for neurotypical participants, no known psychiatric disorders.

Informed consent was obtained from all participants. All procedures were approved by the Université Libre de Bruxelles ethics committee in accordance with the Declaration of Helsinki.

Participants' IQ, Autism Spectrum Quotient (AQ) (Baron-Cohen, Wheelwright, et al., 2001), Empathy Quotient (EQ) (Baron-Cohen & Wheelwright, 2004), and Systemising Quotient (SQ-R) (Baron-Cohen, Richler, Bisarya, Gurunathan, & Wheelwright, 2001) were measured. IQ was assessed in-person, unless participants had recent valid scores they could share; the other measures were collected via online questionnaires. Due to participant dropout, data completeness varied: IQ (90%), AQ (87%), EQ (81%), SQ-R (81%). Table 1 reports means and standard deviations for both groups.

As expected, autistic participants had higher AQ and SQ-R scores and lower EQ scores than neurotypical participants. They also had higher IQs (120 vs. 109). While this should be considered when interpreting results, higher IQ has been found to be either independent from pragmatic ability (Antoniou & Katsos, 2017), or associated with improved pragmatic

	M/F	*Age*	*IQ*	*AQ*	*EQ*	*SQ-R*
ASD	14/20	41 (12)	120 (17)	38 (6)	24 (11)	68 (18)
NT	12/21	29 (11)	109 (12)	17 (7)	41 (11)	55 (15)
Diff	n.s.	$< .001$	.004	$< .001$	$< .001$	.009

Table 1: Participants' demographic features. Standard deviations between brackets. M/F: Number of males and females. IQ: Intelligence quotient. AQ: Autism spectrum quotient. EQ: Empathy quotient. SQ-R: Systemising quotient. Diff: Significance of the difference, as measured with chi-squared tests (for gender) or *t*-tests.

ability (Mazzaggio, Foppolo, Job, & Surian, 2021; Schaeken et al., 2018). Hence, the higher IQ of autistic participants adds a conservative bias against the hypothesis that they behave less pragmatically than neurotypical participants.

Materials and procedure

Participants saw 30 displays containing 100 red or black circles arranged on a grid (Fig. 1). The proportion of red circles varied randomly across trials. Participants were instructed to complete the sentence frame '— are red' ('— est/sont rouge·s' in French). They were explicitly instructed not to use numbers, percentages, or fractions. Otherwise, participants were free to provide any description they saw fit.

The experiment was hosted online on PCIbex Farm (Zehr & Schwarz, 2018), and participants completed it remotely. Exp. 1 is accessible at https://farm.pcibex.net/r/mbqJru/.

Data treatment

Responses were preprocessed by lowercasing, removing diacritics, and eliminating redundant phrases (e.g., 'I see that…' or '… of the circles'). Responses that were intuitively synonymous in truth conditions and conditions of use were manually coded together. Table 2 lists the 16 quantifying expression types analysed and the most frequent tokens for each. English translations are used for the readers' convenience in the main text.

Results

At first glance, autistic and neurotypical participants produced strikingly similar responses. To move beyond impressionistic observation, however, we conducted a series of exploratory analyses.

Our first analysis focused on the length of participants' descriptions. One possibility is that autistic participants might favour greater precision, resulting in longer expressions. Indeed, their responses were, on average, slightly longer (24 characters) than those of neurotypical participants (20 characters). To test whether this difference was reliable, we fit a Bayesian hierarchical regression model predicting the natural logarithm of description length. Group (autistic vs. neurotypical) was entered as a fixed effect, while random

Type	*Tokens*	
about half	environ la moitié	a peu près la moitié
all	tous	
almost all	presque tous	la quasi totalité
	presque la totalité	quasiment tous
few	peu	
fewer than half	moins de la moitié	
half	la moitié	
hardly any	presque aucun	
majority	la majorité	une majorité
many	beaucoup	
minority	la minorité	une minorité
more than half	plus de la moitié	
most	la plupart	
none	aucun	
some	quelques	une partie
vast majority	la grande majorité	une grande majorité
very few	très peu	

Table 2: Quantifying expressions included in the analysis and frequently produced tokens ($n \geq 10$).

intercepts were included for both participants and items. The model was implemented in the 'brms' package (Bürkner, 2017) in R (R Core Team, 2021). Four Markov chains were run for 4,000 iterations each (2,000 warmup, 2,000 sampling), yielding a total of 8,000 posterior draws. We used weakly informative priors: Normal(0, 1) for regression coefficients, and Student-t(3, 0, 2.5) for random effect standard deviations. Convergence was confirmed for all parameters ($\hat{R} < 1.01$, with large effective sample sizes).

The full model specification and parameters can be found in Table 3 in Appendix A. The posterior mean for the autistic–neurotypical contrast was −0.08, with a 95% credible interval ranging from −0.32 to 0.15. Thus, while the point estimate suggests a tendency for autistic participants to produce somewhat longer descriptions, the credible interval includes zero, which leaves the evidence inconclusive. In other words, if such a difference exists, it is small and uncertain.

As a second exploratory analysis, we asked whether autistic participants tended to be more repetitive in their responses, given the commonly reported preference for sameness in autism. To address this, we counted the number of unique quantifying expressions each participant produced across the 30 trials. Autistic participants averaged 20 unique responses, compared to 21 among neurotypical participants. We modelled these counts using a Bayesian Poisson regression analysis with population as a predictor. Again, we ran four chains for 4,000 iterations each, obtaining 8,000 posterior samples, and we placed

weakly informative Normal(0, 2) priors on the regression coefficients. Model diagnostics indicated convergence ($\hat{R} < 1.01$).

The full model specification and parameters can be found in Table 4 in Appendix A. The posterior mean for the group effect was 0.06 (95% CI [−0.04, 0.17]), suggesting that any difference in repetitiveness between groups is minimal at best.

For the following analyses, we focused on a subset of 16 quantifying expressions (Table 2). This decision was guided by a pretest with 24 neurotypical, French-speaking participants, in which we identified the most frequently produced expressions. We retained the 15 most frequently produced quantifying expressions from the pretest, along with the expressions 'all' and 'none', which are both highly salient despite their relatively lower frequency. One additional candidate expression, 'plusieurs' ('several'), was excluded because it occurred in less than 1% of responses in the actual experiment. Together, the selected 16 expressions covered 66% of all responses.

The main motivation for the pretest was to determine which quantifying expressions to test in Exp. 2, which measures participants' intuitions about monotonicity. Since all experiments were conducted in parallel (to limit participant drop-out and optimise the efficient use of valuable participant testing time), we could not wait to analyse the results of Exp. 1 before deciding on which quantifying expressions to test in Exp. 2. As a consequence, the selection of quantifying expressions is based upon patterns of use in neurotypical

participants. We realise that this introduces a bias in that neurotypical behaviour is taken as normative. At the same time, we believe the effect of this bias is limited, since there is a robust correlation between the frequencies of use of different quantifying expressions across the two populations (r = .90, $t(156)$ = 25.7, $p < .001$). Correspondingly, the amount of data that was ultimately included in the analysis was about the same for autistic (68%) and neurotypical (64%) participants.

There were two additional motivations for restricting our attention to 16 quantifying expressions. First, for many of the quantifying expressions that fell outside of the sample, we had extremely few data points, which would not allow the model to reliably estimate the corresponding thresholds. Second, for the quantifying expressions outside of the sample, we did not obtain participants' monotonicity judgements in Exp. 2. Thus, we could not determine whether, according to participants, the quantifying expressions imposed an upper or lower bound (or both) on the intersection set size.

As shown in Fig. 2B, participants' use of these expressions revealed clear effects of pragmatic reasoning. For example, the quantifier 'some' was almost exclusively used in contexts with between 10 and 40 red circles, despite its logical meaning being satisfied by any nonzero number. Similarly, the expressions 'more than half' and 'fewer than half' clustered around proportions close to 50%, rather than being applied across the full logical

range. These patterns confirm that both groups used quantifying expressions in ways that are sensitive to informativeness.

Examining the overall distribution of expressions (Fig. 2A), a few subtle group differences emerged. Whereas neurotypical participants preferred the exact form ‘half’, autistic participants favored the approximate expression ‘about half’. Autistic participants were also somewhat less likely to produce the vague quantifying expressions ‘many’ and ‘very few’. Importantly, however, this was not a general avoidance of vagueness: autistic participants used a variety of other vague expressions (e.g., ‘hardly any’, ‘few’, ‘about half’, ‘almost all’) at least as often as their neurotypical peers. With respect to how the expressions were applied, the clearest divergence concerned ‘many’, which autistic participants used in a less restrictive manner. Still, this observation is tentative, as autistic participants produced only 12 tokens of ‘many’ in total.

To formally assess whether the production distributions differed between groups, we fit a set of Bayesian hierarchical regression models predicting the number of red circles (i.e., intersection set size) from the quantifying expression used. All models included random intercepts for participants. The baseline model included only quantifying expression as a predictor. We then compared this model against two alternatives: one including a main effect of group (autistic or neurotypical), and another including the interaction between

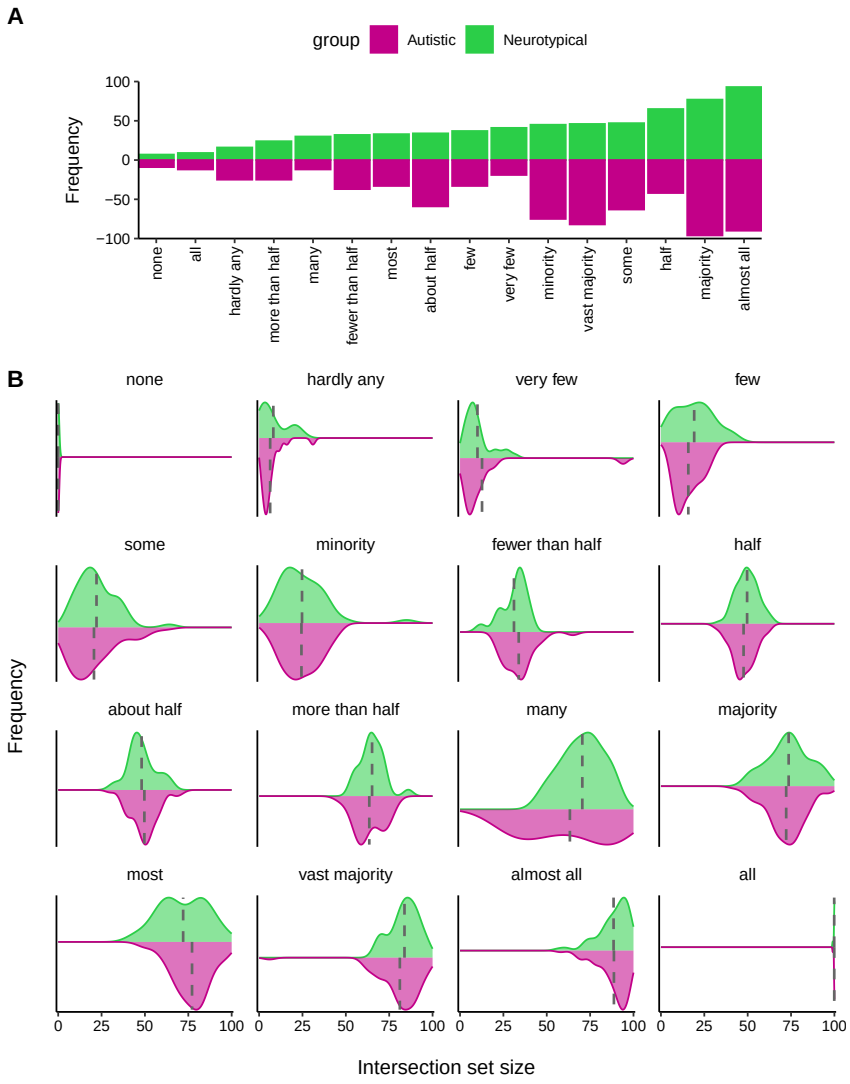

Figure 2: **A**: Overall frequencies of the quantifying expressions that were included in the analyses. **B**: Kernel density plot of the frequencies of these quantifying expressions for different intersection set sizes. Grey dotted lines indicate the mean intersection set size in which the expression was produced.

group and quantifying expression. Model comparison was performed using approximate leave-one-out cross-validation (LOO).

The baseline model provided the best predictive fit. Adding a main effect of group did not improve fit ($\Delta ELPD$ = −0.4, SE = 0.6). Including the interaction actually worsened fit ($\Delta ELPD$ = −9.4, SE = 5.2). Thus, there was no evidence for an interaction effect and little support for a main effect of population. This confirms the impression that the means of the production distributions were very similar for autistic and neurotypical participants.

Taken together, these findings suggest that autistic and neurotypical participants use quantifying expressions in remarkably similar ways, both in terms of the expressions they produce and the contexts in which they use them. Any differences that do exist are minor and localised. Before we turn to the computational modelling of "pragmaticality", we first introduce Exp. 2 and Exp. 3, which provide the necessary parameters for the lexicon and the approximate number system (ANS).

Exp. 2: Monotonicity

Participants

Half of the participants from Exp. 1 were invited to also participate in Exp. 2, while the other half participated in Exp. 3. In total, 25 participants took part in Exp. 2. Of these, 13 had a prior clinical diagnosis of ASD (5 females, mean age: 41.0 years, SD: 10.8, range:

23–58). The remaining 12 participants did not have such a diagnosis (7 females, mean age: 28.5 years, SD: 11.5, range: 18–54). Other demographic characteristics, including IQ, AQ, EQ, and SQ-R, closely resembled those in Exp. 1 and are therefore not repeated here.

The decision to divide participants between Exps. 2 and 3 was pragmatic. Given limited access to participants and the time constraints of testing, each individual could only be asked to complete two experiments. As a result, the sample sizes for both follow-up experiments are relatively small. At the same time, the results of Exp. 2 match our own intuitive judgments, and those of Exp. 3 align with previous findings, providing at least some reassurance about their reliability. Nevertheless, because of the restricted sample sizes, the findings from these experiments should be interpreted with caution.

Materials and procedure

The experiment tested participants' judgments of validity for two types of inferences: upward and downward. Upward arguments involved reasoning from a set to its superset (e.g., from 'eat salmon' to 'eat fish'), while downward arguments involved reasoning from a set to its subset (e.g., from 'eat fish' to 'eat salmon').

(8) *Upward argument*

Premise: Some of the people ate salmon.

Conclusion: Some of the people ate fish.

(9) *Downward argument*

Premise: Some of the people ate fish.

Conclusion: Some of the people ate salmon.

For each participant, the 16 quantifying expressions from Exp. 1 were paired with two predicates forming a set-subset relation.[2] These predicates were adapted from the materials of van Tiel et al. (2021, Exp. 2; see their Supporting Information), who had pretested them to ensure that the set-subset relation was transparently understood (e.g., that 'eat salmon' necessarily denotes a subset of 'eat fish'). In some cases, we modified predicates to suit the European context of our study; for example, replacing references to U.S. cities (e.g., 'come from Chicago') with European ones (e.g., 'come from Paris').

Each quantifying expression appeared in both an upward and a downward argument, with the same pair of predicates presented in reversed order. Participants judged whether each argument was valid; that is, whether the conclusion had to be true if the premise was true, or equivalently, whether it was impossible for the premise to be true and the conclusion false. Responses were given by clicking 'yes' or 'no'. To familiarise participants with the notion of validity, two annotated examples were presented at the beginning of

[2]In this task, we also included the expression 'plusieurs' (≈ 'several'). As noted earlier, this expression was not retained for analysis, since it had been produced too rarely in Exp. 1.

the experiment, and the definition of validity was displayed as a reminder under each argument.

Exp. 2 can be accessed at https://farm.pcibex.net/p/AdbqkC/.

Results

For every participant and quantifying expression, we classified responses into four types: (i) *Upward-only* (accepting only the upward argument), (ii) *Downward-only* (accepting only the downward argument), (iii) *Neither*, or (iv) *Both*. Fig. 3 shows the frequency of these response types across quantifying expressions.

Upward-only responses indicate that a quantifying expression was treated as monotone increasing; *Downward-only* responses as monotone decreasing; *Neither* suggests a non-monotone interpretation (neither increasing nor decreasing). *Both* responses likely reflect errors or misunderstandings of the task. Such responses were rare (19 out of 400 total responses, or 4.75%) and were excluded from subsequent analyses.

Although the task employed deductive validity instructions, our aim was not to assess deductive competence in the abstract. Rather, we use participants' validity judgements to infer how the quantifying expressions are interpreted in this task. The deductive format ensures a clear criterion (impossibility of premise true and conclusion false), but the analysis is descriptive with respect to which semantic profile participants appear to

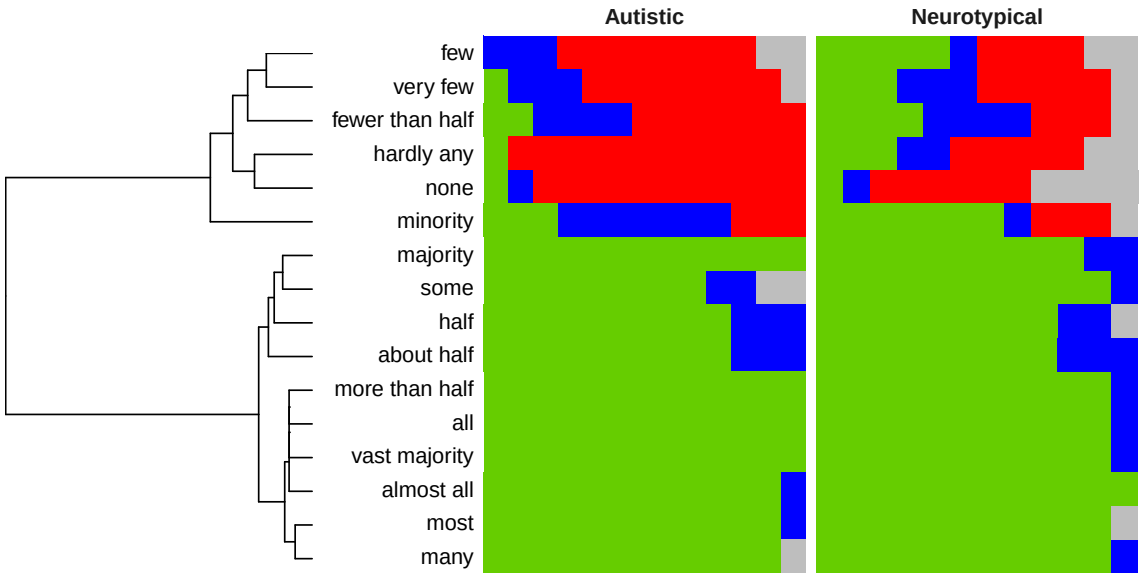

Figure 3: Response frequencies in Exp. 2. Bars show the proportion of different response types for each quantifying expression. On the left, a dendrogram displays the outcome of a cluster analysis grouping expressions by similarity in response patterns.

assign to each expression. In this sense, the task provides behavioural evidence about interpretation under deductive instructions, rather than a normative test of logical ability.

To supply the model with monotonicity classifications, we applied a hierarchical cluster analysis to determine which expressions patterned with 'all' (unambiguously monotone increasing) and which with 'none' (unambiguously monotone decreasing). This data-driven clustering was preferred over modal-response classification alone in order to control for participants' general bias towards accepting upward inferences.

For clustering, we coded *Upward-only* responses as 1, *Neither* as 2, and *Downward-only* as 3. A Euclidean distance matrix was then computed, and Ward's method was applied for hierarchical clustering. Analyses were initially run separately for autistic and neurotypical participants, but as the results were identical, groups were collapsed for the final classification.

The clustering results (Fig. 3) showed that 'a minority', 'few', 'fewer than half', 'hardly any', and 'very few' patterned with 'none', and were therefore classified as monotone decreasing. All remaining quantifying expressions clustered with 'all', and were classified as monotone increasing.

Most expressions were straightforward to classify and confirmed our expert intuitions, especially given the general participant bias towards accepting upward inferences. An exception was 'a minority', which yielded roughly equal proportions of *Upward-only*,

Downward-only, and *Neither* responses. This variability mirrors our own less certain intuitions about this expression. Still, we classified it as monotone decreasing, since it patterned with 'none' and, in our judgement, typically licenses inferences from a set to its subset (e.g., from 'A minority of people ordered fish' to 'A minority of people ordered salmon'), which—if correct—indicates that it does not lexically impose a lower bound, since otherwise such inferences would be invalid.

We emphasise that there is no one-to-one mapping between natural-language quantifying expressions and logical quantifiers, and that interpretations may vary across individuals. Ideally, monotonicity would be estimated at the participant level and propagated into the computational model. However, given the limited number of responses per participant and per expression, such an approach would be statistically underpowered and would substantially increase model complexity. We therefore adopt a population-level classification as a pragmatic modelling choice, while noting that some expressions (most notably 'a minority') may exhibit genuine interpretive variability.

This modelling decision does not imply that logical monotonicity itself is determined by majority opinion. The monotonicity of a formally specified quantifier is a logical property. However, identifying which quantifier a natural-language expression denotes is, in many cases, an empirical question. For example, whether 'almost all' encodes only a lower bound or both a lower and upper bound has consequences for its monotonicity

behaviour. The strong preference for upward inferences observed here suggests that any upper bound associated with this expression is pragmatic rather than semantic. Thus, while logical monotonicity is a normative notion, determining which formal operator best captures an expression's meaning requires empirical evidence.

Importantly, our classification concerns the bounds that are lexically encoded. Many expressions may give rise to additional lower- or upper-bounding inferences at the level of interpretation. For example, 'some' is often understood as 'some but not all', and 'fewer than half' may suggest 'fewer than half, but at least one'. Within our modelling framework, such additional bounds are treated as pragmatic enrichments arising from competition with alternatives, rather than as part of the lexical semantic specification.

Participants' monotonicity judgements largely accorded with this assumption. If, for instance, 'some' lexically encoded both a lower and an upper bound, participants should systematically reject both upward and downward inferences; however, such non-monotone response patterns were relatively infrequent (approximately 16% for 'some'). Introducing a two-bounded lexical representation would therefore risk conflating pragmatic strengthening with encoded meaning and would add parameters not independently warranted by the present dataset.

At the same time, the data reveal greater variability for negative quantifiers such as 'few' or 'hardly any' than for positive ones such as 'some' or 'almost all'. Reasoning about

negative information is known to be cognitively demanding (Clark & Chase, 1973), and this processing asymmetry likely contributes to the observed dispersion in responses. More generally, uncertainty about the precise monotonicity profile of certain expressions could be incorporated directly into the computational model, either at the population level or, ideally, at the level of individual participants. However, the additional parameters and data requirements of such an approach would substantially increase model complexity beyond what the present dataset can reliably support. Our classification should therefore be understood as an instrumental modelling decision, which provides a tractable approximation while leaving open the possibility of more fine-grained modelling in future work.

As an exploratory analysis, we examined the extent to which participants' responses aligned with the monotonicity classification derived above. For this purpose, we provisionally treat that classification as a target profile, but we recognise that it simplifies underlying variability. Because of its marked heterogeneity, 'a minority' was excluded from this analysis.

For each quantifying expression classified as monotone increasing, we measured how often participants accepted upward inferences and rejected downward ones; for expressions classified as monotone decreasing, the scoring was reversed. We refer to this

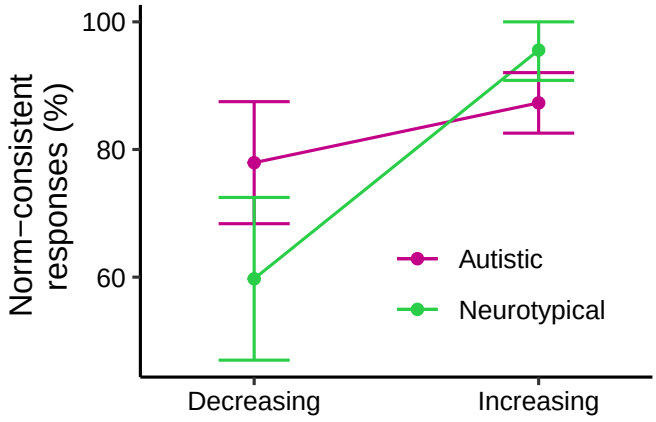

Figure 4: Participants' mean rate of norm-consistent responses for monotonicity inferences. Error bars represent 95% confidence intervals.

measure as *norm-consistency*, defined relative to the assumed monotonicity profile. Fig. 4 displays mean norm-consistency by group and type of quantifying expression.

To analyse these data, we fit a Bayesian hierarchical logistic regression model with norm-consistency (consistent/inconsistent) as the outcome. Predictors were population (autistic vs. neurotypical), quantifier type (monotone increasing vs. decreasing), and their interaction, with IQ included as a covariate. Random intercepts were included for both participants and quantifying expressions. The model was fit with four chains of 4,000 iterations each, using weakly informative Normal(0, 2) priors for regression coefficients and Student-$t(3, 0, 2.5)$ priors for random-effect standard deviations. The model converged ($\hat{R} < 1.01$; large effective sample sizes).

The full model specification and parameters can be found in Table 5 in Appendix A. The posterior estimates indicated that autistic participants were overall more norm-consistent than neurotypical participants (β = -1.14, 95% CI [−2.23, −0.14]). Norm-consistency was higher for monotone increasing than for monotone decreasing expressions (β = 1.27, 95% CI [0.46, 2.30]). Importantly, there was also a positive interaction (β = 1.40, 95% CI [0.42, 2.40]), showing that autistic participants were especially norm-consistent on monotone decreasing expressions. IQ had no detectable effect (β = 0.01, 95% CI [−0.01, 0.04]).

In summary, autistic participants' judgements aligned more closely with our empirically-driven monotonicity classification, particularly for monotone decreasing quantifiers. Importantly, this does not imply superior logical competence per se, but rather a closer correspondence between their response patterns and the monotonicity profile inferred from the present data. Whether this reflects differences in logical competence, pragmatic enrichment, sensitivity to alternative interpretations, or other processing factors remains an open question for future research.

Exp. 3: Estimation

Participants

Half of the participants from Exp. 1 were invited to take part in Exp. 3, while the other half completed Exp. 2. In total, 28 participants participated in Exp. 3. Nine of these had a

prior clinical diagnosis of ASD (6 females, mean age: 39.0 years, SD: 8.6, range: 25–52), and the remaining 18 did not (13 females, mean age: 30.9 years, SD: 11.6, range: 19–60). Other demographic measures (IQ, AQ, EQ, and SQ-R) were highly similar to those reported in Exp. 1 and are therefore not repeated here.

Materials and procedure

The displays in this experiment were identical to those used in Exp. 1 (Fig. 1). Each display consisted of 100 circles, coloured either red or black. The number of red circles was randomly determined for each trial, with the constraint that no display was shown twice to the same participant. Each participant completed 25 such trials.

Unlike Exp. 1, where participants described the displays using quantifying expressions, here they were asked to give exact numerical estimates. A horizontal slider ranging from 0 to 100 was presented below each display, and participants adjusted the slider to indicate the number of red circles. The corresponding number was displayed beneath the slider.

Exp. 3 can be accessed at https://farm.pcibex.net/p/vQMFmD/.

Results

Fig. 5 shows participants' numerical estimates against the true intersection set size. To test whether autistic and neurotypical participants differed in their estimation accuracy,

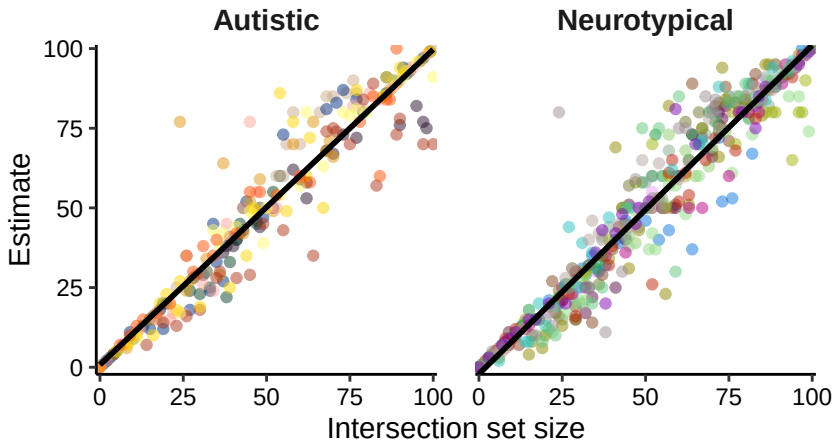

Figure 5: Participants' estimates of the intersection set size. Each colour represents a different participant.

we modelled the standardised error of each estimate as a function of population (autistic vs. neurotypical), with random intercepts for both participants and items.

We fit a Bayesian linear mixed-effects regression with a Gaussian likelihood, implemented with four chains of 4,000 iterations each (2,000 warmup, 2,000 sampling). Weakly informative priors were used: Normal(0, 1) for fixed effects, and Student-t(3, 0, 1) for both group-level and residual standard deviations. Posterior distributions were summarised with means and 95% credible intervals. Model diagnostics indicated convergence ($\hat{R} < 1.00$; large effective sample sizes).

The full model specification and parameters can be found in Table 6 in Appendix A. The model output revealed no credible group difference (β = 0.03, 95% CI [−0.23, 0.30]).

Because estimation accuracy was comparable across populations, we proceeded under the assumption of a single Weber fraction shared by both groups.

Weber's fraction was calculated as follows. Let $d_i = \langle \text{Act}_i, \text{Est}_i \rangle$ denote the i-th datapoint, where Act_i is the actual intersection set size and Est_i is the participant's estimate. We determined the parameter pair $\langle w, \epsilon \rangle$ that maximised the Laplace-smoothed log-likelihood:

$$\log P(D \mid w, \epsilon) = \sum_i \log P(\text{Est}_i \mid \text{Act}_i, w, \epsilon), \text{ where}$$

$$P(\text{Est}_i \mid \text{Act}_i, w, \epsilon) \propto P_{\text{ANS}}(\text{Est}_i \mid \text{Act}_i, w) + \epsilon$$

Here, $P_{\text{ANS}}(\text{Est}_i \mid \text{Act}_i, w)$ was defined as in the model specification provided in the section 'A model for the production of quantifying expressions'.

This analysis yielded a Weber fraction of 0.266 for the approximate number system (ANS) component of the model. For comparison, van Tiel et al. (2021) reported a considerably larger Weber fraction of 0.576. The discrepancy likely reflects differences in display complexity: whereas our participants estimated proportions out of 100 circles, van Tiel and colleagues used more demanding displays containing 432 circles.

Model fitting

We fit the computational model to the production data from Exp. 1 using Stan (Stan Development Team, 2024). Each model was run with four chains of 5,000 warmup iterations

followed by 3,000 sampling iterations, yielding a total of 8,000 posterior draws. We ensured sample quality based on visual inspection, convergence based on $\hat{R}$ (Gelman & Rubin, 1992) (judged by $\hat{R} < 1.1$ for all parameters), and other diagnostics such as absence of divergent transitions. Overall, the model captured the production data well. The correlation between predicted and observed frequencies was high for both groups: $r = 0.89$ for autistic participants and $r = 0.88$ for neurotypical participants.

The model included several free parameters. The first set of parameters concerned the *thresholds* associated with quantifying expressions. Each expression specifies either a lower or upper bound, depending on its monotonicity, as determined in Exp. 2. These thresholds were treated as free parameters. Formally, thresholds θ_m were drawn from a scaled Beta distribution:

$$\frac{\theta_m}{100} \propto \text{Beta}(\omega_m, \kappa_m)$$

Here, ω_m represents the modal threshold value and κ_m the concentration. For quantifying expressions with standard logical interpretations (e.g., 'all', 'none', 'more than half', 'fewer than half'), priors were centered on their textbook meanings. For vague or context-dependent expressions (e.g., 'few', 'many', 'almost all'), priors were chosen based on linguistic intuition. Most priors were weakly informative ($\kappa = 10$), though we

imposed stronger priors on ‘all’ and ‘none’ ($\kappa = 40$) to encode their logical precision. Fig. 6 illustrates the prior and posterior distributions for each threshold.

For most expressions, the posterior thresholds aligned well with our prior expectations. For example, ‘none’ centered near 0, ‘some’ near 2, ‘more than half’ near 56, and ‘all’ near 100. A notable exception was ‘a minority’, which received a surprisingly high upper bound (83). This result likely reflects a small number of atypical productions in contexts with many red circles. Posterior distributions were generally narrow, suggesting that the data strongly constrained the threshold estimates.

To assess robustness, we re-estimated the model after halving and doubling the prior concentration parameters κ, while holding the prior means fixed. Across these specifications, the qualitative pattern of posterior threshold estimates and the group comparison for λ remained unchanged, although doubling κ led to reduced bulk effective sample sizes for some parameters. This indicates that the main conclusions are not driven by the specific choice of prior concentration.

The second set of parameters concerned the *salience* of each expression. Salience determines how likely an expression is to be chosen, independent of informativeness. Because salience enters the model as a multiplicative factor, only relative values matter. We therefore represented the vector of salience values, s, as a normalised probability vector with a symmetric Dirichlet(1) prior:

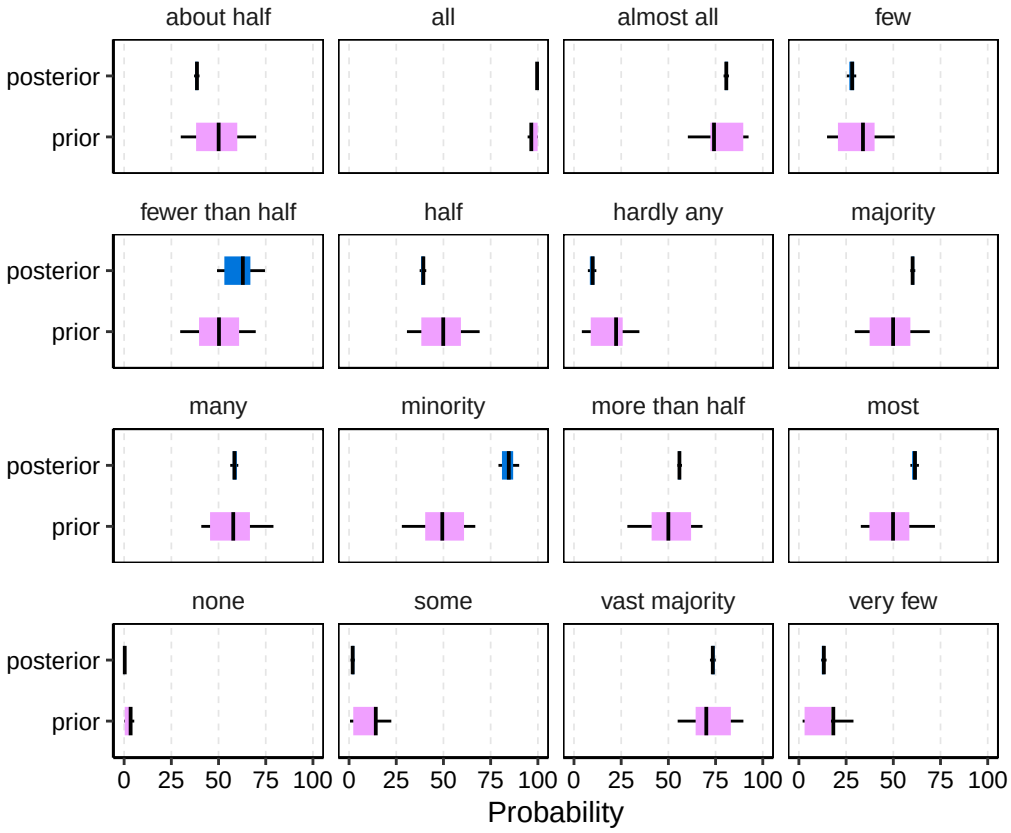

Figure 6: Prior and posterior distributions over threshold parameters. Vertical bars show medians; coloured areas show 50% highest density regions; black horizontal lines show 80% highest density regions.

$$s \approx \text{Dirichlet}(\overrightarrow{1})$$

Here, $\overrightarrow{1}$ is a vector of 16 1s (one for each quantifier in the analysis). Posterior estimates of salience values are shown in Fig. 7. These values mirrored the empirical production frequencies (Fig. 2), revealing a robust bias in favour of monotone increasing quantifying expressions (those that specify a lower bound, such as 'some' or 'most') over monotone decreasing quantifying expressions (those that specify an upper bound, such as 'few' or 'fewer than half').

Finally, and most crucially, the model included a free "pragmaticality" parameter λ. This parameter governs how strongly a speaker prefers informative messages; that is, messages from which a literal listener is likely to infer the intended intersection set size. In contrast to thresholds and salience, which were estimated across all participants, we fit separate λ parameters for autistic and neurotypical groups. This allowed us to examine potential group differences in the weighing of pragmatic optimisation.

Both λ parameters were assigned weakly informative lognormal priors (location log(2), scale 1). Posterior estimates (Fig. 8) yielded mean values of 2.33 (95% CI [2.01, 2.68]) for neurotypical participants and 1.88 (95% CI [1.62, 2.17]) for autistic participants. The posterior distribution of the group difference ($\lambda_{\text{NT}} - \lambda_{\text{ASD}}$) was concentrated above zero

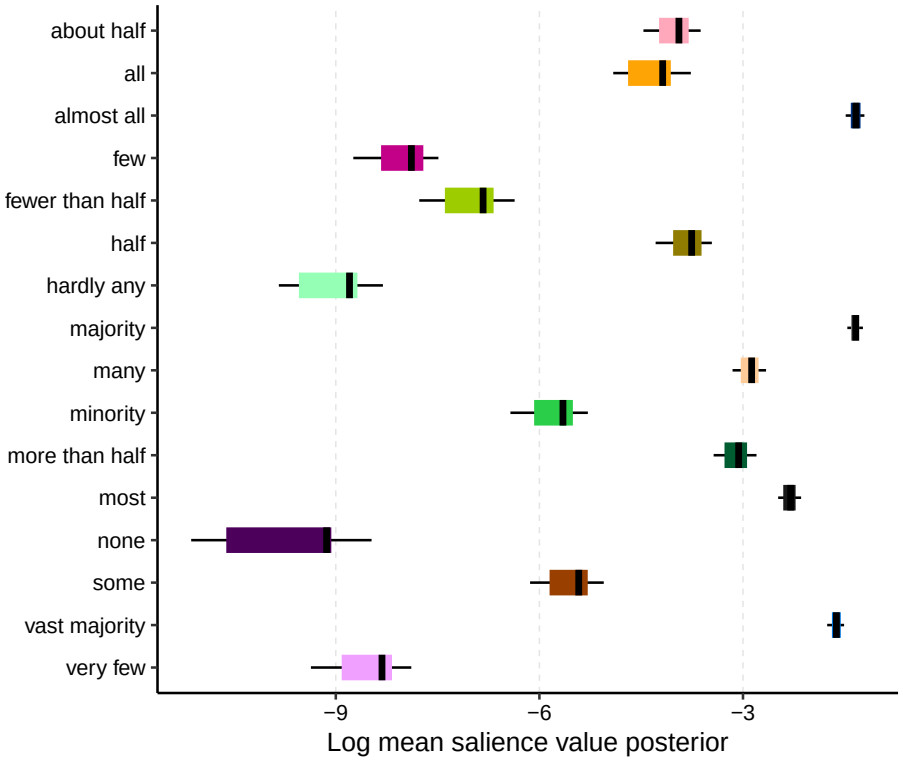

Figure 7: Posterior distributions over log salience values. Vertical bars are means; coloured areas show 50% highest density regions; black horizontal lines show 80% highest density regions.

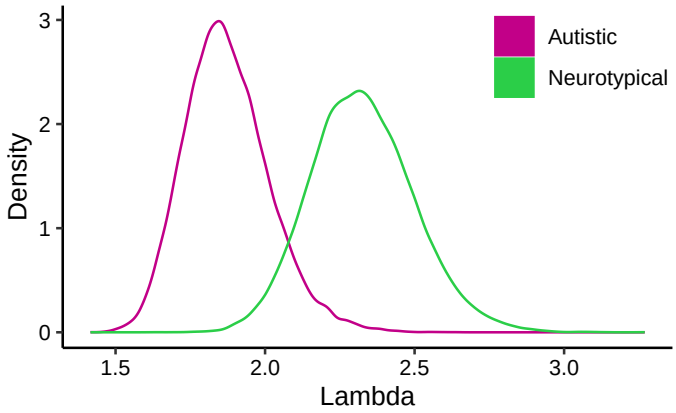

Figure 8: Posterior distributions over pragmaticality parameter λ for autistic and neurotypical participants.

(posterior probability > 0.999), which indicates consistent directional evidence for higher λ values in the neurotypical group.

At the same time, the magnitude of the difference was modest (mean difference $\approx$ 0.45). Thus, while the posterior distribution supports a directional group difference within the model, the effect size is small.

To put this difference in perspective, consider a simplified situation in which the total set size is ten, and the speaker has only six messages at their disposal to describe the intersection set size I, with the following lexical entries: 'none' means $I \leq 0$; 'few' means $I \leq 3$; 'some' means $I \geq 1$; 'most' means $I \geq 6$, 'almost all' means $I \geq 8$; and 'all' means $I \geq 10$. Fig. 9 shows production probabilities for four types of speakers in this scenario: a nearly literal speaker (who produces any true message), an optimally pragmatic speaker

(who almost always selects the most informative message), and two intermediate speakers whose λ values match the posterior means for our autistic and neurotypical groups.

As the figure shows, both estimated speaker models behave far more like optimally pragmatic speakers than literal ones. The group difference is subtle. For example, when all circles are red, the autistic speaker model produces 'all' with probability .84, compared to .90 for the neurotypical speaker model.

Importantly, this modelling result should be interpreted alongside the distributional analysis reported in the Results section of Exp. 1, which showed no improvement in predictive fit when group was added as a factor. Thus, at the level of observed production distributions, group differences are minimal. The computational model analysis instead suggests a modest difference in the softmax parameter λ, which governs how strongly speakers' production probabilities track differences in communicative utility.

In sum, the computational modelling indicates a modest and model-dependent group difference. Autistic participants are slightly less likely to select informative messages, but their behaviour remains strongly pragmatic overall.

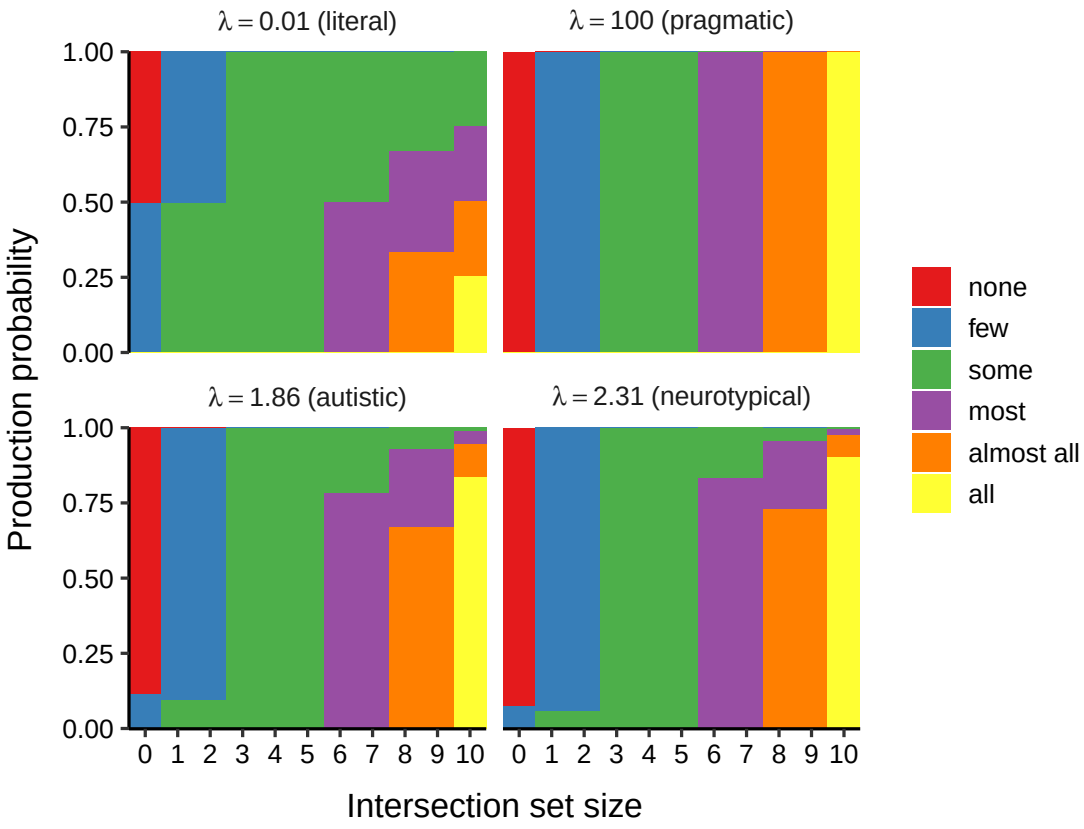

Figure 9: Production probabilities associated with different values of the "pragmaticality" parameter λ in a simplified scenario with eleven possible intersection set sizes, and six possible messages.

General discussion

Summary of findings

There are usually many different ways of truthfully describing a situation. A pragmatically competent speaker is able to arbitrate between these alternatives and select a message that is *informative*—that is, one that is likely to receive the intended interpretation (e.g., Grice, 1975).

Autism is frequently characterised by pragmatic difficulties. Accordingly, one might expect that autistic people have difficulties navigating the space of alternatives in order to select an utterance that is both truthful and informative. In this study, we leveraged a recent computational model for the production of quantifying expressions to test whether autistic participants were indeed less likely to produce optimally informative descriptions (van Tiel et al., 2021).

In line with our predictions, we found that the model's "pragmaticality" parameter was estimated to be higher for neurotypical than for autistic participants. At the same time, the difference was relatively modest, pointing to a difference in degree rather than kind.

Theoretical implications

Our results suggest that autistic people do not face substantial problems with pragmatic reasoning about informativeness. Like their neurotypical peers, they associate quantifying expressions with pragmatically justified ranges of situations that typically form only a subset of the logically possible ones. For instance, both autistic and neurotypical participants used 'some' to describe situations in which roughly 20–40% of the circles were red, even though it is logically appropriate whenever there is at least one red circle.

These findings support previous research that challenges the view of autism as involving a global pragmatic deficit (e.g., Brock, Norbury, Einav, & Nation, 2008; Kissine et al., 2015; Norbury, 2005). They also highlight the heterogeneity of pragmatic phenomena. Pragmatics comprises a diverse range of linguistic behaviours that differ along multiple dimensions, making it unlikely that a single cognitive mechanism underpins them all, or that autism entails a uniform impairment across them (e.g., Geurts et al., 2019; Jary et al., 2025; Kissine, 2021; Marrocchini, 2023).

One important dimension along which pragmatic phemomena vary is their degree of *conventionality*, i.e., the extent to which repeated usage has linked particular expressions with particular interpretations. To illustrate, compare the following two metaphorical sentences:

(10) a. Boris Johnson is a cat with nine lives.

b. Boris Johnson is a tin of baked beans.

The first sentence is easily understood because ‘a cat with nine lives’ has a conventionalised figurative meaning. The second sentence is more opaque, requiring pragmatic inference from context (here, it followed the remark that British voters knew they were not getting “a tin of caviar” when they elected Johnson) and background knowledge.

Given autism’s association with a preference for sameness, autistic people may experience greater difficulty with pragmatic phenomena that demand flexible, context-dependent interpretation, while showing relatively preserved ability with conventionalised forms (e.g., Deliens, Papastamou, Ruytenbeek, Geelhand, & Kissine, 2018; Kissine, 2012; Koch, van Langen, Bašnáková, & Stolk, 2026; Mangnus et al., 2024; Wadge, Brewer, Bird, Toni, & Stolk, 2019). Entrenched metaphors such as ‘a cat with nine lives’ are one example of conventionalised pragmatics; the focalised use of quantifiers is another. For instance, because speakers typically use ‘some’ to describe proportions around 20–40%, this mapping may become encoded in one’s mental representation of the expression. If so, this could explain why autistic participants in our study used quantifiers in ways closely resembling their neurotypical peers.

This account leads to a testable prediction: autistic individuals may show more difficulty with quantifying expressions that rely heavily on contextual information for interpretation. Consider:

(11) Many people liked your talk.

The interpretation of 'many' depends on various contextual factors, such as whether the speaker is a Ph.D. student or a senior professor, whether the talk was controversial, or the size of the audience (Cummins & Franke, 2021; Lorson, Cummins, & Rohde, 2021; Schöller & Franke, 2016). It would be informative to test whether autistic participants flexibly adjust their interpretations of such quantifying expressions to contextual variation.

Methodological contribution

Methodologically, our study illustrates how theoretically motivated probabilistic models can uncover subtle population-level differences in language use (see also Ramotowska, Steinert-Threlkeld, van Maanen, & Szymanik, 2024; Sarafoglou et al., 2024). Previous studies reported no differences between autistic and neurotypical participants in interpreting 'some' as excluding 'all', suggesting comparable use of quantifying expressions (e.g., Chevallier et al., 2010; Pijnacker et al., 2009). By contrast, our probabilistic modelling

revealed subtle group-level differences in latent pragmatic ability. We encourage further use of such models to revisit cases where traditional analyses have yielded null results.

Limitations

Our study has several limitations. First, the two groups were not perfectly matched, with autistic participants having higher IQs than neurotypical participants. While prior developmental work suggests no negative association—and in some cases a positive link—between IQ and pragmatic ability (Antoniou & Katsos, 2017; Mazzaggio et al., 2021; Schaeken et al., 2018), little is known about this relationship in adults. We were unable to disentangle the relative contributions of IQ and autism diagnosis, leaving this as a question for future research.

Second, the ecological validity of our production task can be questioned. Participants were instructed not to provide numerical responses and to use verbal quantifying expressions instead. Although we attempted to make verbal quantification natural by presenting displays with many circles, in order to make exact enumeration effortful, the question format ('How many circles were red?') may nevertheless have made numerical descriptions salient. This creates a potential tension between the conversational framing of the task and the restriction to verbal expressions (see Erev & Cohen, 1990; Juanchich & Sirota, 1990, for related work on preferences for verbal vs. numerical quantifying expressions).

Crucially, however, in the RSA framework pragmatic optimality is always defined relative to a specified set of alternatives and a communicative objective. By restricting the response space to verbal quantifying expressions, we define a constrained but well-specified decision problem. The model therefore evaluates optimality within that alternative set, rather than with respect to all logically possible utterances. It remains an open question whether similar patterns would emerge when competition between verbal and numerical expressions is modelled directly or when tasks perhaps more naturally elicit verbal quantification.

A third limitation is that we modelled behaviour at the population rather than participant level. This means that we do not explicitly capture individual variability in either pragmatic reasoning or quantifying expression interpretation within groups. In particular, monotonicity profiles were derived at the population level, and individual-level semantic heterogeneity was not modelled directly (cf. Ramotowska et al., 2024). Given the sparse number of observations per participant and per expression, reliable participant-specific estimation was not feasible. Future work could address this by collecting denser individual data. Nevertheless, the present population-level results provide a useful counterpoint to earlier reports of population-level null effects (e.g., Chevallier et al., 2010; Pijnacker et al., 2009).

A fourth limitation concerns the preselection of quantifying expressions. Because the experiments were conducted in parallel and Exp. 2 required a fixed set of items, the expression set was determined using a pretest with neurotypical participants. This could be seen as introducing bias by taking neurotypical usage patterns as a starting point. However, pragmatic optimality in our analyses is defined relative to a model-based criterion rather than conformity to group averages. Moreover, expression frequencies were highly correlated across groups (r = .90), and the proportion of included data was comparable, suggesting that the selection procedure did not systematically disadvantage autistic participants. Future work could nevertheless replicate the design using independently specified expression sets or fully data-driven inclusion criteria.

Finally, our study used a non-interactive setting. While this may have reduced social pressure and cognitive load (particularly beneficial for autistic participants), it prevented us from examining how interaction partners shape expression use. Prior work shows that the presence, and especially the neurotype, of interlocutors influences autistic communication (Davis & Crompton, 2021; Milton, 2012). Mixed-neurotype dyads often have less successful interactions, likely due to divergent communicative strategies (Heasman & Gillespie, 2019; Williams, Wharton, & Jagoe, 2021). Future research should therefore extend these questions to interactive contexts.

Conclusion

Effective use of quantifying expressions requires reasoning about the listener's likely interpretation, a skill often presumed to be impaired in autism. While earlier studies reported no group differences in the interpretation of quantifying expressions (e.g., Chevallier et al., 2010; Pijnacker et al., 2009), our computational modelling provided a finer-grained perspective. We found that autistic participants were less likely than neurotypical participants to select optimally informative messages, though the difference was subtle. This demonstrates the value of bespoke probabilistic modelling in uncovering nuanced group differences and provides a quantitative estimate of the effect of autism on pragmatic ability.

Data availability statement

The anonymised data and analysis files associated with this article can be accessed via the following link:

https://osf.io/gfxde/overview?view_only=6acf16bd2c144ecf84e3160477fc66f5

References

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

Antoniou, K., & Katsos, N. (2017). The effect of childhood multilingualism and bilectalism on implicature understanding. *Applied Psycholinguistics*, *38*, 787–833.

Ashwood, K. L., Gillan, N., Hayward, H., Woodhouse, E., McEwen, F. S., Findon, J., … 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. (2013). *Mindblindness: An essay on autism and theory of mind.* MIT Press.

Baron-Cohen, S., Richler, J., Bisarya, D., Gurunathan, N., & Wheelwright, S. (2001). The systemizing quotient: An investigation of adults with Asperger syndrome or high-functioning autism, and normal sex differences. *Philosophical Transactions of the Royal Society of London B*, *358*, 361–374.

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: An investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, *34*, 163–175.

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.

Bartczak, M. (2017). Processing metaphors in the elderly: Does valence matter? *Psychology*

of Language and Communication, *21*, 352–379.

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

Begeer, S., Rieffe, C., Meerum Terwogt, M., & Stockmann, L. (2003). Theory of mind-based action in children from the autism spectrum. *Journal of Autism and Developmental Disorders*, *33*, 479–487.

Boehm, N. L. (2021, July 13). *Neurodiversity.* Defeating Epilepsy Foundation. Retrieved from https://www.defeatingepilepsy.org/understanding-epilepsy/neurodiversity-2 (Accessed: 2025-10-21)

Brock, J., Norbury, C. F., Einav, S., & Nation, K. (2008). Do individuals with autism process words in context? Evidence from language-mediated eye-movements. *Cognition*, *108*, 896–904.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software*, *80*, 1–28.

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., & Chase, W. G. (1973). On the process of comparing sentences against

pictures. *Cognitive Psychology*, *3*, 472–517.

Cummins, C., & Franke, M. (2021). Rational interpretation of numerical quantity in argumentative contexts. *Frontiers in Communication*, *6*, 662027.

Davis, R., & Crompton, C. J. (2021). What do new findings about social interaction in autistic adults mean for neurodevelopmental research? *Perspectives on Psychological Science*, *16*, 649–653.

Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, *9*, 519–540.

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

Deliens, G., Papastamou, F., Ruytenbeek, N., Geelhand, P., & Kissine, M. (2018). Selective pragmatic impairment in autism spectrum disorder: Indirect requests vs irony. *Journal of Autism and Developmental Disorders*, *48*, 2938–2952.

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.

Frank, M. C., & Goodman, N. D. (2012). Predicting language reasoning in language games.

Science, *336*, 998.

Geelhand, P., Papastamou, F., Belenger, M., Clin, E., Hickman, L., Keating, C. T., & Sowden, S. (2023). Autism-related language preferences of French-speaking autistic adults: An online survey. *Autism in Adulthood*, *5*, 275–288.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, *7*, 457–472.

Gernsbacher, M. A., & Yergeau, M. (2020). Empirical failures of the claim that autistic people lack a theory of mind. *Archives of Scientific Psychology*, *7*, 102–118.

Geurts, B., Kissine, M., & van Tiel, B. (2019). Pragmatic reasoning in autism. In K. Morsanyi & R. Byrne (Eds.), *Thinking, reasoning, and decision-making in autism* (pp. 113–134). Routledge.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*, 173–184.

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

Heasman, B., & Gillespie, A. (2019). Neurodivergent intersubjectivity: Distinctive features of how autistic people create shared understanding. *Autism*, *23*, 910–921.

Hochstein, L., Bale, A., & Barner, D. (2017). Scalar implicature in the absence of epistemic reasoning? The case of autism spectrum disorder. *Language Learning and Development*, *14*, 224–240.

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

Horn, L. R. (1989). *A natural history of negation.* Chicago University Press.

Jary, M., Martín-González, I., Vicente, A., & Castroviejo, E. (2025). Performance of autistic adults on conversational implicatures: A comparison of material and behavioural inferences. *Journal of Communication Disorders*, *115*, 106509.

Juanchich, M., & Sirota, M. (1990). Verbal versus numerical probabilities: Efficiency, biases and the preference paradox. *Psychological Research*, *45*, 2325–2338.

Khorsheed, A., & van Tiel, B. (2024). Why second-language speakers sometimes, but not always, derive scalar inferences like first-language speakers: Effects of task demands. *Language Acquisition.* (Advance online publication)

Kim, S. H., Paul, R., Tager-Flusberg, H., & Lord, C. (2014). Language and communication in autism. In F. R. Volkmar, S. J. Rogers, R. Paul, & K. A. Pelphrey (Eds.), *Handbook of autism and pervasive developmental disorders* (4th ed., pp. 230–262). Wiley.

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

Kissine, M. (2021). Autism, constructionism, and nativism. *Language*, *97*, e139–e160.

Kissine, M., Cano-Chervel, J., Carlier, S., De Brabanter, P., Ducenne, L., Pairon, M.-C., … Leybaert, J. (2015). Children with autism understand indirect speech acts: Evidence from a semi-structured act-out task. *PLoS ONE*, *10*, e0142191.

Koch, S. B. J., van Langen, J., Bašnáková, J., & Stolk, A. (2026). Partner-dependent communication without dynamic adaptation in autism. *Autism*, *30*, 736–747.

Krifka, M. (2002). Be brief and vague! And how bidirectional optimality theory allows for verbosity and precision. In D. Restle & D. Zaefferer (Eds.), *Studies in structure and change: A festschrift for Theo Vennemann* (pp. 439–458). Mouton de Gruyter.

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

Leslie, A. M., & Thaiss, L. (1992). Domain specificity in conceptual development: Neuropsychological evidence from autism. *Cognition*, *43*, 225–251.

Lorson, A., Cummins, C., & Rohde, H. (2021). Strategic use of uncertainty expressions. *Frontiers in Communication*, *6*, 635156.

Lorson, A., Macuch-Silva, V., Hart, C., & Winter, B. (2024). Gesture size affects numerical estimates in quantifier comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* (Advance online publication)

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.

Mangnus, M., Koch, S. B. J., Cai, K., Greidanus Romaneli, M., Hagoort, P., Bašnáková, J., & Stolk, A. (2024). Preserved spontaneous mentalizing amid reduced intersubject variability in autism during a movie narrative. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *10*, 1057–1066.

Marrocchini, E. (2023). Impairment or difference? The case of theory of mind abilities and pragmatic competence in the autism spectrum. *Applied Psycholinguistics*, *44*, 365–383.

Masi, A., DeMayo, M. M., Glozier, N., & Guastella, A. J. (2017). An overview of autism spectrum disorder, heterogeneity and treatment options. *Neuroscience Bulletin*, *33*, 183–193.

Mazzaggio, G., Foppolo, F., Job, R., & Surian, L. (2021). Ad-hoc and scalar implicatures in

children with autism spectrum disorder. *Journal of Communication Disorders*, *90*, 106089.

Mazzaggio, G., & Surian, L. (2018). A diminished propensity to compute scalar implicatures is linked to autistic traits. *Acta Linguistica Academica*, *65*, 651–668.

Milton, D. E. (2012). On the ontological status of autism: The 'double empathy problem'. *Disability & Society*, *27*, 883–887.

Montague, R. (1973). The proper treatment of quantification in ordinary English. In H. J. J. Hintikka, J. M. E. Moravcsik, & P. Suppes (Eds.), *Approaches to natural language* (pp. 221–242). Reidel.

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

Moxey, L., & Sanford, A. J. (1993). *Communicating quantities.* Lawrence Erlbaum.

Newstead, S. E., & Griggs, R. A. (1984). Fuzzy quantifiers as an explanation of set inclusion performance. *Psychological Research*, *46*, 377–388.

Norbury, C. F. (2005). The relationship between theory of mind and metaphor: Evidence from children with language impairment and autistic spectrum disorder. *British Journal of Developmental Psychology*, *23*, 383–399.

Noveck, I. A. (2018). *Experimental pragmatics: The making of a cognitive science.* Cambridge University Press.

Parker, E. (2023, April 20). *Is epilepsy neurodivergent?* Goally. Retrieved from https://getgoally.com/blog/is-epilepsy-neurodivergent/ (Accessed: 2025-10-21)

Pastor-Cerezuela, G., Tordera Yllescas, J. C., González-Sala, F., Montagut-Asunción, M., & Fernández-Andrés, M.-I. (2018). Comprehension of generalized conversational implicatures by children with and without autism spectrum disorder. *Frontiers in Psychology*, *9*, 272.

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.

R Core Team. (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Ramotowska, S. (2022). *Quantifying quantifier representations: Experimental studies, computational modeling, and individual differences* (Unpublished doctoral dissertation). University of Amsterdam, The Netherlands.

Ramotowska, S., Steinert-Threlkeld, S., van Maanen, L., & Szymanik, J. (2024). *Most* quantifiers have *many* meanings. *Psychonomic Bulletin and Review*, *31*, 2692–2703.

Sarafoglou, A., Giacobello, A., Godmann, H. R., Johnson, T., Visser, I., Haaf, J. M., & Szymanik, J. (2024). *A Bayesian framework to study individual differences in semantic representations.* Retrieved from https://osf.io/preprints/osf/hvgb7 (OSF Preprints)

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

Schaeken, W., Van Haeren, M., & Bambini, V. (2018). The understanding of scalar implicatures in children with autism spectrum disorder: Dichotomized responses to violations of informativeness. *Frontiers in Psychology*, *9*, 1266.

Schlotterbeck, F., Ramatowska, S., van Maanen, L., & Szymanik, J. (2020). Representational complexity and pragmatics cause the monotonicity effect. In *Proceedings of the 43nd Annual Meeting of the Cognitive Science Society* (pp. 3398–3404).

Schöller, A., & Franke, M. (2016). How many are many? Exploring semantic theories with data-driven computational models. In N. Bade, P. Berezovskaya, & A. Schöller (Eds.), *Proceedings of Sinn und Bedeutung* (pp. 622–639).

Schöller, A., & Franke, M. (2017). Semantic values as latent parameters: Testing a fixed threshold hypothesis for cardinal readings of *few* & *many*. *Linguistic Vanguard*, *3*, 159–219.

Stan Development Team. (2024). *RStan: the R interface to Stan.* Retrieved from https://

mc-stan.org/ (R package version 2.32.6)

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.

Tager-Flusberg, H., Paul, R., & Lord, C. (2005). Language and communication in autism. In F. Volkmar, R. Paul, A. Klin, & D. Cohen (Eds.), *Handbook on autism and pervasive developmental disorders* (pp. 335–364). Wiley.

van Tiel, B. (2014). Embedded scalars and typicality. *Journal of Semantics*, *31*, 147–177.

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*, 1373–1403.

van Tiel, B., Sauerland, U., & Franke, M. (2022). Meaning and use in the expression of estimative probability. *Open Mind*, *6*, 250–263.

Vicente, A., & Falkum, I. L. (2023). Accounting for the preference for literal meanings in autism spectrum conditions. *Mind and Language*, *38*, 119–140.

Volden, J., Coolican, J., Garon, N., White, J., & Bryson, S. (2009). Brief report: Pragmatic language in autism spectrum disorder: Relationships to measures of ability and disability. *Journal of Autism and Developmental Disorders*, *39*, 388–393.

Wadge, H., Brewer, R., Bird, G., Toni, I., & Stolk, A. (2019). Communicative misalignment in autism spectrum disorder. *Cortex*, *115*, 15–26.

Williams, G. L., Wharton, T., & Jagoe, C. (2021). Mutual (mis)understanding: Reframing autistic pragmatic "impairments" using relevance theory. *Frontiers in Psychology*, *12*, 616664.

World Health Organisation. (2022). *International classification of diseases (ICD)-11.* Retrieved from https://icd.who.int/en

Zehr, J., & Schwarz, F. (2018). *PennController for Internet Based Experiments (IBEX).* Retrieved from https://doi.org/10.17605/OSF.IO/MD832

Appendix A: Model output

Exp. 1: Comparison of description length

log(length) ∼ group + (1 | participant) + (1 | item)

Parameter	**Estimate**	**SE**	**95% CI**
Fixed effects			
Intercept	2.83	0.08	[2.67, 2.99]
Group (neurotypical)	-0.08	0.11	[-0.31, 0.14]
Random effects (SDs)			
Participant (Intercept)	0.47	0.04	[0.40, 0.57]
Item (Intercept)	0.16	0.02	[0.12, 0.19]
Residual variance			
σ	0.51	0.01	[0.49, 0.52]

Table 3: Estimates of posterior means with 95% credible intervals.

Exp. 1: Comparison of unique responses

unique-responses ∼ group

Parameter	**Estimate**	**SE**	**95% CI**
Fixed effects			
Intercept	3.00	0.04	[2.92, 3.07]
Group (neurotypical)	0.06	0.05	[-0.04, 0.17]

Table 4: Estimates of posterior means with 95% credible intervals.

Exp. 2: Monotonicity accuracy comparison

correct $\sim$ group * type + IQ + (1 | participant) + (1 | quantifier)

Parameter	**Estimate**	**SE**	**95% CI**
Fixed effects			
Intercept	-0.00	1.84	[-3.70, 3.67]
Group (neurotypical)	-1.15	0.51	[-2.15, -0.12]
Type (Increasing)	1.38	0.47	[0.46, 2.35]
IQ	0.01	0.01	[-0.01, 0.04]
Group $*$ Type	1.41	0.52	[0.39, 2.45]
Random effects (SDs)			
Participant (Intercept)	0.82	0.26	[0.39, 1.39]
Quantifier (Intercept)	0.46	0.26	[0.04, 1.39]

Table 5: Estimates of posterior means with 95% credible intervals.

Exp. 3: Numerosity estimation accuracy comparison

error ∼ group + (1 | participant) + (1 | item)

Parameter	**Estimate**	**SE**	**95% CI**
Fixed effects			
Intercept	-0.02	0.11	[-0.25, 0.20]
Group (neurotypical)	0.03	0.14	[-0.24, 0.30]
Type (Increasing)	1.38	0.47	[0.46, 2.35]
Random effects (SDs)			
Participant (Intercept)	0.28	0.06	[0.18, 0.41]
Item (Intercept)	0.40	0.05	[0.30, 0.51]

Table 6: Estimates of posterior means with 95% credible intervals.