Supplementary Materials

1. Supplemental analysis details

Data, code, and analysis scripts are available in our OSF repository: https://osf.io/jmp7u

1.1 Experiment 1

The final regressions used for each of the three analyses were:

Proximal ~ Perspective + (1 + Perspective | Participant) + (1 + Perspective | Language)

Proximal ~ Closer + Location + (1 + Location | Participant) + (1 | Language)

Marked distal ~ Farther + Location + (1 + Location | Participant) + (1 | Language)

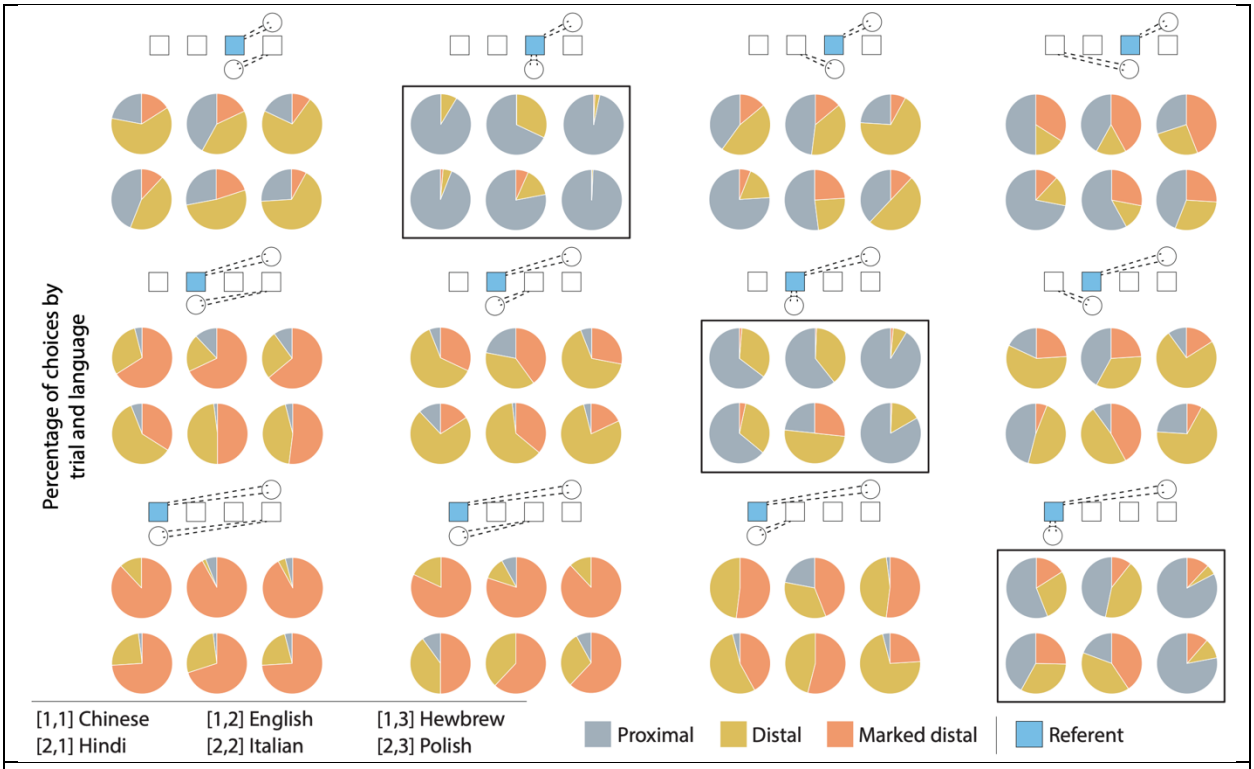

Figure S1. Full results from Experiment 1. Each collection of six pie charts shows the results for one of the 12 events we test. The schematic above shows the relevant event, with the speaker at the top, the listener at the bottom, and the target referent highlighted in blue. Events surrounded by a black box indicate joint attention. Each row of events shows a case where the referent is matched but the listener's attention changes. Thus, each collection of six pie charts reveals the

consistency of choices across languages. Each row of events then shows how these patterns of demonstrative choice change as a function of listener attention, controlling for all physical aspects of the event.

1.2 Experiments 2a and 2b

Analyses followed a parallel logic to the ones from Experiment 1, using a separate analysis for Experiments 2a and 2b. In Experiment 2b, analyses focused on medial demonstrative choice because that indicates proximity to the listener (see Supplementary Text for prediction details). The first regressions were binomial mixed-effects models predicting proximal (Exp 2a) and medial (Exp 2b) demonstrative choice as a function of whether the perspectives were aligned or misaligned, using the maximal random effects structures that converged:

Exp 2a: Proximal ~ Perspective + (1 + Perspective | Participant) + (1 | Language)

Exp 2b: Medial ~ Perspective + (1 + Perspective | Participant) + (1 | Language)

The second regressions were parallel to the one in Experiment 1, and focused on misaligned trials and tested proximal choice as a function of whether the listener should look closer, using the maximal random effects structure that converged. Both regressions had identical form:

Exps 2a and 2b: Proximal ~ Closer + Location + (1 + Location | Participant)

The final regressions were also parallel to the one in Experiment 1, and focused on misaligned trials and tested distal choice as a function of whether the listener should look farther, using the maximal random effects structure that converged:

Exp 2a: Distal ~ Farther + Location + (1 + Location | Participant)

Exp 2b: Distal ~ Farther + Location + (1 + Location | Participant) + (1 | Language)

1.3 Experiment 2c

Analysis approach was the same as Experiments 2a-2b.

The maximal converging model to test increased use of the proximal when perspectives are aligned was

Proximal ~ Perspective + (1 + Perspective | Subject)

To test whether the proximal was also used more when attention should be pulled, the converging maximal model was

Proximal ~ Closer + Position + (1 + Position | Subject)

Conversely, the maximal regression to test increased use of the distal when attention should be pushed was

$$\text{Distal} \sim \text{Farther} + \text{Position} + (1 + \text{Position} \mid \text{Subject})$$

Note however, that the Proximal and Distal variables are mirror images (because in this study participants only used the proximal or the distal). Similarly, the Closer and Farther variables are also mirror images (since these regressions only use misaligned trials, all trials must either required pushing or pulling attention). Therefore, the two regressions above are mathematically equivalent.

Figure S2. Trial-by-trial results of Experiment 2c with Mandinka speakers. Each barplot represent distribution of demonstrative choices for the event shown directly above it. Vertical black lines represent 95% confidence intervals. Barplots surrounded by a black rectangle indicate events in joint attention.

1.4 Experiment 4

In Task 3, the following options were available for selection when asked about what situations they use 'este', 'ese' or 'aquel':

ESTE (proximal)
- For objects close to you
- For objects close to the other person
- When the other person was looking further than the object you wanted.

ESE (medial)
- For objects far from you
- For objects close to the other person
- When the other person was looking at the object you wanted.

AQUEL (distal)
- For objects far from you
- For objects far from the other person
- When the person was looking closer than the object you wanted.

Analysis approach
Open-box responses in Task 1 were coded by two independent coders, who were given the following coding dimensions:

- Speaker distance
- Listener distance
- Speaker pointing
- Speaker reach
- Listener reach
- Speaker visual attention
- Listener visual attention
- Example of demonstrative use

The Task 3 regression was as follows:

Listener-att ~ Att-sensitivity + (1 + Att-sensitivity | Demonstrative) + (1 | Participant)

2. Estimating attention correction effects

Figure 5 in the main manuscript presents a visualization for how different demonstratives are used to manipulate attention. These estimates were obtained through logistic regressions trained

to predict people's use of demonstratives through least squares error minimization (but not used as statistical tests for significance).

For each demonstrative, we ran a logistic regression with demonstrative choice as the dependent variable (coded as 1 for target demonstrative and 0 for all other options). Independent variables were:

- Offset: distance between target referent and listener attention, with positive values indicating referent is farther away from the speaker relative to listener attention, and negative values indicating referent is closer to speaker relative to listener attention.

- Aligned: binary variable indicating whether the agents were in joint attention (i.e., listener already looking at the referent).

- Position: Distance of the referent from the speaker, numerically coded.

This regression therefore allowed us to estimate the effects of joint attention and attention correction on demonstrative choice while controlling for physical location. Figure 5 shows the regression estimates for the offset predictor (on x axis) and the estimate for the aligned predictor (y axis).

3. Supplementary theoretical background on polysemy predictions

Our main analyses focused on three predictions. First, proximal demonstratives are used to pull attention. Second, distal demonstratives are used to push attention. Finally, some demonstratives are polysemous, encoding both joint attention and attention correction. The first two predictions were inspired by past work arguing that demonstratives are used to establish joint attention, with the difference being that previous analyses assumed that the attention correction occurred at the pragmatic level (Enfield, 2003), with demonstrative semantics encoding only a position in space (Diessel, 2013).

We further predicted that the proximal demonstrative would be polysemous in Experiment 1, based on a debate in the literature. While some typologists argued that demonstratives indicate spatial distance to a point of reference (Diessel, 2014), some psycholinguists argued that demonstratives indicate psychological proximity, rather than a purely spatial one (Peeters & Özyürek, 2016). In this second view, spatial representations are scaled to the shared space created by the location of the interlocutors. Building on this debate, we hypothesized that in languages with two-way systems, the proximal demonstrative is polysemous between the spatial meaning (gloss: "the one close to me") and the mentalistic meaning (gloss: "the one in joint attention").

Based on this, we reasoned that three-way demonstrative systems should also have a polysemous demonstrative form that encodes both joint attention and attention correction. Critically, however, three-way demonstrative systems are not equivalent to a two-way system plus a third medial demonstrative (despite the misleading re-use of the terms 'proximal' and 'distal' for both system types), as their semantics are known to be different (Diessel, 2013; 2014). Therefore, past literature was insufficient to predict which demonstrative form (the proximal, which indicates proximity to the speaker, or the medial, which indicates proximity to the listener) might be

polysemous in a three-way system. We reasoned that, in principle, it is possible that both the proximal and medial could be used to mark joint attention.

However, our experimental design in Experiment 2 was matched to Experiment 1, and this design only included trials where the referent was in front of the speaker. This made us unable to test for the potential polysemous meaning of the proximal in 3-way systems, because the spatial meaning of that demonstrative is believed to encode proximity to speaker and distance to listener (gloss: "the one close to me and far from you" (Diessel, 2013; 2014). Therefore, our experimental design did not include trials that could elicit the spatial meaning of the proximal. By contrast, the spatial meaning of the medial is believed to encode distance from speaker and proximity to listener (gloss: "the one far from me and close to you" (Diessel, 2013; 2014), and our design therefore included trials what would elicit this meaning, which then enabled us to test if the selection of this demonstrative form varied when attention was misaligned.

Therefore, our omission of a polysemy test for the proximal demonstrative in 3-way systems does not imply lack of a potential prediction, but only an inability to test it with our current design. Critically, however, the evidence that the medial encodes both spatial meanings and attention correction is sufficient to validate our main hypothesis that mentalistic representations are also in the grammar of three-way demonstrative systems.

4. Computational framework

All model code is available in our OSF repository. Below we explain the conceptual structure.

4.1 Lexicon and word semantics

We define a Lexicon L as a collection of referential expressions. In our experimental context, the lexicons can consist of two-way systems (e.g., $L=\{$this one, that one$\}$), two-way systems with a marked distal (e.g., $L=\{$this one, that one, that one over there$\}$), or three-way systems (e.g., $L=\{$este, ese, aquel$\}$).

Each word is associated with a semantic mapping that assigns a probability to each possible referent. In our experimental context, there are four possible positions in space, so each semantic mapping is a probability distribution over the four possible referents. For instance, a spatial semantics of the proximal demonstrative (e.g., "this one") would assign the highest probability to the position closest to the speaker and decrease the probability for positions farther away.

We use S to indicate a semantic mapping where $S(D,R)$ is the probability that demonstrative D identifies referent R. Throughout we consider three types of semantics:

4.1.1 Spatial semantics. Spatial semantics are represented through Beta distributions discretized over the table's four spatial regions. The proximal and distal spatial semantics were represented as mirror images: The proximal through a $\mathrm{Beta}(1,X)$ distribution and the distal through a $\mathrm{Beta}(X,1)$ distribution where $X>1$. The market distal was modeled as a $\mathrm{Beta}(Y,1)$ distribution where $Y>X$ (i.e., creating an even stronger bias for mapping this referent to farther-away regions from the speaker, and with a sharper decay).

In three-way demonstrative systems, the use of the medial demonstrative to indicate intermediate regions can be implemented in two ways. A first possibility is through a semantic mapping that favors middle regions (e.g., Beta(2,2)). A second possibility is through a pragmatic process, where the medial is concentrated in the middle region, due to the strong association of the proximal to close regions and the distal to far regions.

4.1.2 Joint attention semantics. Joint attention semantics are represented as a simple distribution that assigns probability 1 to objects in joint attention and 0 otherwise. Critically, words encoding joint attention are polysemous and so these joint attention semantics must be combined with other semantic mappings. Rather than averaging the semantics, we give our model the joint attention semantics as a separate concept. For instance, in two-way demonstrative systems, our model has access to two conceptually different words that sounds the same. The first has the spatial semantic mapping and the second has the joint attention semantic. Thus, although the model ultimately provides a single probability for the utterance, it internally distinguishes between the *sense* (i.e., which of the two semantic mappings) it intends to use.

4.1.3 Attention correction semantics. Attention correction semantics are represented as uniform distributions over the direction of attention. That is, the semantics of 'look closer' are represented as a uniform distribution over all referents closer to the speaker, and the semantics of 'look further' are represented as a uniform distribution over all referents farther away from the speaker. This semantic mapping can be averaged with the spatial semantic (such that the resulting meaning is sensitive to both the spatial location of the referent and the listener's attention).

4.2 Pragmatic model (RSA)

4.2.1 Prior beliefs. Once pragmatic reasoning is available, the speaker can consider the listener's mental states. In our experimental paradigm this means that the speaker can represent the listener's referential expectations (i.e., "what does the listener think I'm going to talk about?") and use them to decide what to say. Formally, this requires specifying a prior probability distribution over reference.

We consider three possible prior beliefs about the listener's referential expectations:

- Basic prior: The basic prior consists of a uniform distribution over all referents.

- Sharp_Att_Prior: The sharp attention prior captures a situation where the listener is looking at the object she thinks will be the referent, and believes that all other objects are equally (un)likely to be the referent. This prior is implemented through a utility vector that assigns value 1 to all possible referents and value 2 to the one the listener is looking at. This vector is then turned into a probability distribution via softmaxing:

$$p(x) \propto e^{U(x)/\tau}$$

  Where τ is the softmax parameter.

- Graded_Att_Prior: The graded attention prior is conceptually similar to the sharp attention prior. However, rather than placing a uniform distribution over all objects the

agent is not looking at, objects closer to where the listener is looking at are more likely to be the referent (that is, an object one position away from their attention is more likely to be the referent than an object two positions away from their attention).

We implement this prior through a utility vector where the utility of each referent is given by

$$U(x) = 4 - |\Delta(a, x)|$$

Where a is the listener's attention (such that $U(x) = 4$ for the object the listener is looking at, and decays linearly as a function of distance. This utility vector is then transformed into a probability distribution via softmaxing.

4.2.2 RSA. Our pragmatic model is a standard RSA implementation, where agents engage in recursive social reasoning to infer referential intentions. We begin with an L_0 listener that selects a referent R by probability matching the semantics of the demonstrative D. That is,

$P_{L0}(R|D) = S(D,R)$

Next, an S_1 speaker selects demonstrative D by

$P_{S1}(D|R) \propto P_{L0}(D|R)$

The pragmatic listener L_1, then reasons about S_1 with a prior over the speaker's referential expectations, such that

$P_{L1}(R|D) \propto P_{S1}(D|R)P_{S1}(R)$

Note that while $P_{S1}(R)$ is a prior over what speaker S_1 will talk about, this prior is in the listener's mind and may be incorrect.

Finally, the pragmatic speaker S_2 uses the pragmatic listener when deciding which demonstrative to use, given by

$P_{S2}(D|R) \propto P_{L1}(R|D)$.

4.3 Model space

Given the three basic types of semantic mappings available, we consider the following three models, all equipped with RSA pragmatics.

- Basic: All demonstratives have purely spatial semantics
- JA: All demonstratives have spatial semantics. In addition, one of the demonstratives can also encode joint attention.

- JA_AC: All demonstratives have spatial semantics. However, the proximal also has a 'pulling attention' semantics and the distal has a 'pushing attention' semantics. In addition, one of the demonstratives encodes joint attention (the proximal in two-way systems, and the medial in three-way systems).

Combined with three possible priors, this creates a space of nine possible models for each demonstrative system. For simplicity and clarity, we evaluated these models in two-way demonstrative systems, this reduces the need to define the two additional semantic mappings for the medial and the marked distal. The model code on our OSF repository is implemented so that these additional demonstratives (or any arbitrary number of additional demonstratives) can be easily integrated.

4.4 Model predictions

The model, due to its probabilistic nature, was run 1000 times to calculate the predicted proportion of demonstrative use.

4.4.1 Qualitative analysis. The main goal of our computational models is to explore what types of semantics can give rise to demonstratives that signal joint attention and attention correction. Our first focus is therefore on the key qualitative patterns we documented experimentally. For this question, quantitative differences in the strength of semantics cannot produce novel quantitative patterns (e.g., how tight a spatial range the proximal has will not affect whether the model spontaneously uses the proximal for attention correction or not). Thus, in exploring this question we set the proximal semantics to a $\mathrm{Beta}(1,4)$ mapping, and the distal semantics to a $\mathrm{Beta}(4,1)$. Qualitative simulations used a softmax parameter $\tau = 1$ (for the two graded priors).

The behavior of all nine models is visualized below in Figure S3.

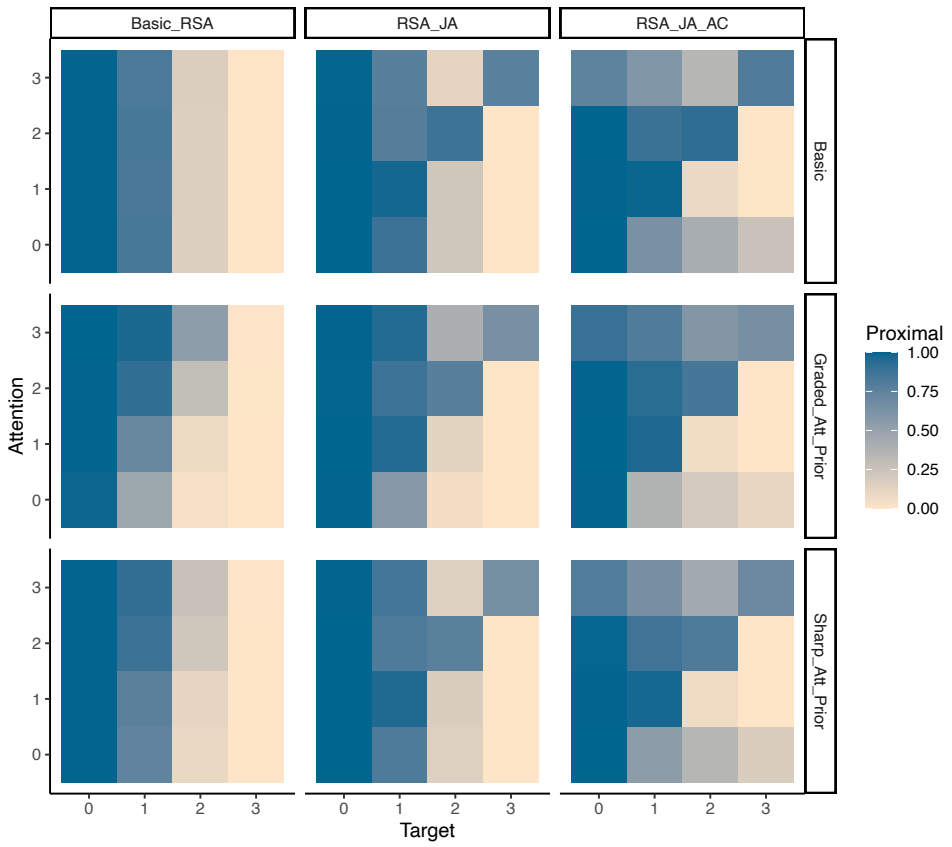

Figure S3. Use of the proximal demonstrative as a function of target location (x axis) and listener attention (y axis). In each plot, joint attention should produce a diagonal pattern of proximal use (the positions where target and listener attention are aligned), and attention correction should produce an asymmetry between the top left quadrant (where attention must be pulled) and the bottom right quadrant (where attention must be pushed). Each column of plots represents a different set of semantics, and each row represents a different prior over referential intentions.

Note also that the Graded_Att_Prior and Sharp_Att_Prior do not produce any differences, confirming that the pragmatic effect of attention correction emerges only from the expectation that the listener might have a false belief (i.e., is looking at the object she expects to be the referent). We additionally confirmed that the qualitative pattern of results is identical when a marked distal is included (setting the marked distal semantics to a `Beta(6,1)` mapping) as can be seen in Figure S4.

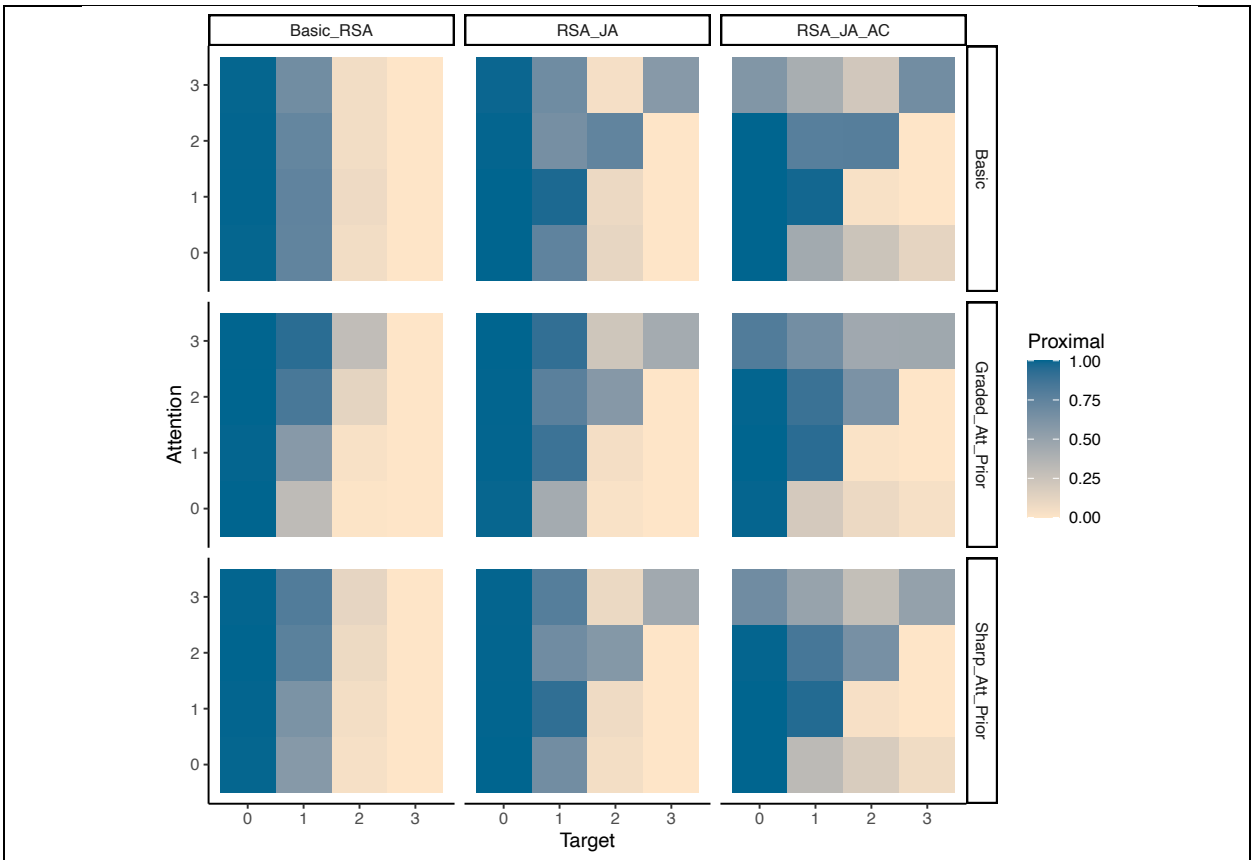

Figure S4. Replication of analyses in Figure S3 with the addition of a marked distal. All qualitative patterns are identical, showing only a reduced use of the proximal (due to the addition of a marked distal that was available to speakers).

4.4.1 Quantitative analysis. Our main text focused on the key models developed to understand the relationship between pragmatics and semantics: Basic_RSA with a Basic prior, RSA_JA with a Basic prior, RSA_JA_AC with a Basic prior, and RSA_JA with Graded_Att_Prior. To compare them directly to participant data, it is possible that the parameter setting could affect the quantitative fit. That is, while the results about which models successfully generate the qualitative phenomena are not affected by the parameter choice, the exact fit to people's demonstrative choices can be affected.

To solve this, the simulation results presented in the main text reflect the model predictions integrated over a range of parameters. Specifically, we introduced a *Semantic Bias* parameter which can consist of any integer in the range [2,10]. This *Semantic Bias* then determined the strength of the spatial bias of each demonstratives. Specifically, the spatial semantics were modeled as a Beta(1, *Semantic Bias*) for the proximal demonstrative and Beta(*Semantic Bias*, 1) for the distal demonstrative (Note that we do not include 1 in the range of semantic biases because this would induce a uniform distribution). We set a Gaussian distribution with mean = 6 and standard deviation 1.5 as the prior over the *Semantic Bias*. This was set so that the mean would match the mean of the range of semantic biases that we tested, and the standard deviation functionally concentrated all the probability mass in the [2,10] range. The rationality parameter

Tau, when applicable (i.e., for the attention-based priors), was tested over the range [0.01,2] in increments of 0.1, and we used a uniform prior over rationality values.

5. References

1. Enfield, N. J. (2003). Demonstratives in space and interaction: Data from Lao speakers and implications for semantic analysis. *Language*, *79*(1), 82-117.

2. Diessel, H. (2013). Where does language come from? Some reflections on the role of deictic gesture and demonstratives in the evolution of language. *Language and Cognition*, *5*(2-3), 239-249.

3. Diessel, H. (2014). Demonstratives, frames of reference, and semantic universals of space. *Language and Linguistics Compass*, *8*(3), 116-132.

4. Peeters, D., & Özyürek, A. (2016). This and that revisited: A social and multimodal approach to spatial demonstratives. *Frontiers in psychology*, *7*,173316.