Plurality in Mandarin vs English: new experimental perspectives

ESPP Conference 2025

Claire RONG
clairerongfr@gmail.com

Based on joint work with Prof. Benjamin Spector

Department of Cognitive Studies, École Normale Supérieure, Paris

September 2, 2025

Outline

Outline

A logical gap

Bare plurals in English most often give rise to a **multiplicity inference**:

(1) The box contains books.

In some environments (e.g. negative sentence), the meaning of bare plurals is **not** the negation of their meaning in simple affirmative sentences:

(2) The box doesn't contain books.

The **logical gap**: situations where the box contains **exactly one book**.
Is **exactly one book** included in the denotation of "books"?

How do we account for the logical gap?

Accounting for the logical gap: hypotheses

Hypothesis 1: bare plurals are inherently ambiguous between the readings *at least one* and *at least two*. > Thought to be wrong (Spector 2007)

Hypothesis 2: one of the two readings is the literal semantic denotation of a bare plural and the other reading is derived from the literal meaning.

Accounting for the logical gap: (main) existing theories

Hypothesis 2: one of the two readings is the literal semantic denotation of a bare plural and the other reading is derived from the literal meaning.

Hypothesis 2 instantiated by 2 classes of approaches:

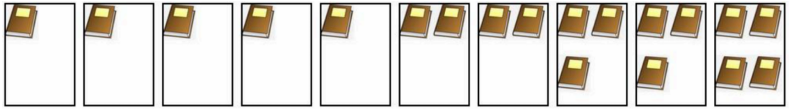
- **Bivalent approaches:** bare plurals have an *at least one* denotation, which gets pragmatically strengthened to *at least two*.
  - Higher-Order implicature (Spector 2007)
  - Zweig(+Ivlieva)'s approach (Zweig 2007, Ivlieva 2020)
- **Trivalent approaches:** bare plurals have truth conditions (*at least two*), falsity conditions (*zero*) and are undefined for *exactly one*.
  - Presuppositional Exhaustification approach (Ahn, Saha, and Sauerland 2020)
  - Homogeneity-based approach (Križ 2017)

Predictions of the different theories

All those theories make **different predictions** on the truth conditions, in **mixed situations**, of a bare plural in the scope of a universal quantifier:

(3) Each box contains books.

Mixed situations: some boxes contain exactly one book and others multiple.



Predictions of the different theories

To describe their predictions, we need to introduce different readings:

Each box contains books.

1. **Literal** reading: every box contains one or more books.
2. **Weak** reading: every box contains one or more books and it is not the case that every box contains exactly one book.
3. **Strong** reading: every box contains several books.

Logical strengths: strong > weak > literal.
Consequence: we **cannot** test **any combination** of readings.

HOI approach	{literal, weak, strong}
Zweig(+Ivlieva)'s approach	{strong}
Presuppositional Exhaustification approach	{literal(#), strong}
Homogeneity-based approach	{literal(#), strong}

Theoretical questions

Core Theoretical Question 1

What are the available readings?

Core Theoretical Question 2

How universal are the mechanisms of plural interpretation?
More specifically, as a case study, what are the available readings in Mandarin, a language with optional number marking?

Outline

Outline

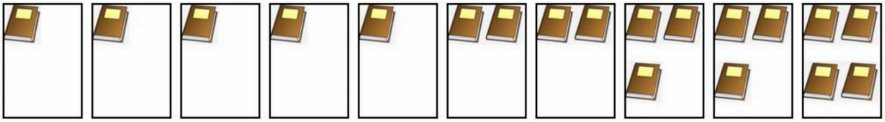
Gradient effects in production: previous literature

Gradient effects: phenomenon where continuous truth-value judgments vary with the distribution within a mixed scenario.

→ **Not predicted by any theory**, but have been observed in previous literature, in both language production (Enguehard 2024) and comprehension (Jiang and Sudo 2023; Stateva, Andreetta, and Stepanov 2016).

Gradient effects in production: our experiment

277 native English speakers recruited on Prolific.
Each participant saw each condition once.



Please observe the picture and complete the sentence below.

Every box contains

Experimental condition: number of boxes containing a unique object.

Each condition instantiated 11 times, with 11 different objects (apples, bikes, birds, books, chairs, flowers, houses, pencils, rabbits, stars, trees).

Production study in English: hypothesis

Hypothesis

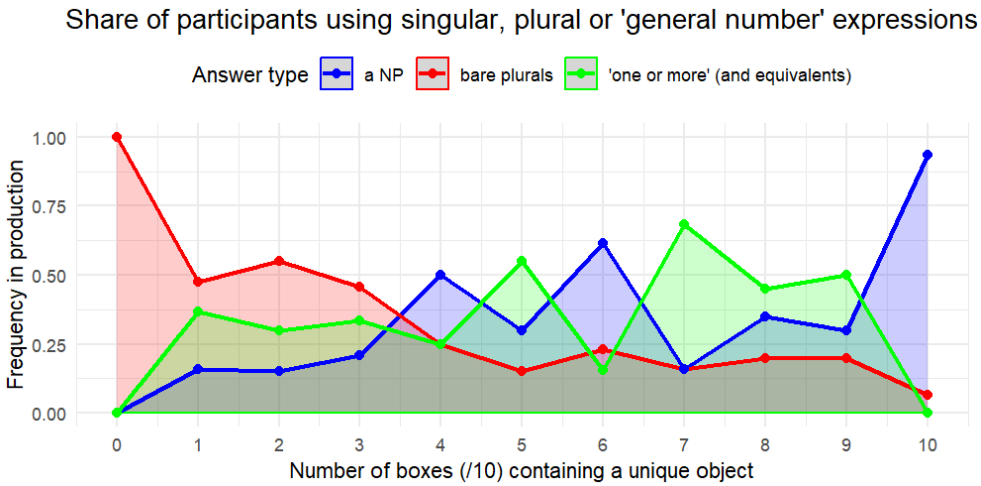
The higher the proportion of boxes containing a unique object, the greater the proportion of participants who will use a singular noun to complete the universally quantified sentence: *Every box contains...* Conversely for bare plurals.

We fit a logistic regression model predicting the log-odds of choosing a singular noun as a function of the number of boxes containing a single object.
The full model was compared with a null model by means of a LRT (likelihood ratio test).

Predictions

$b > 0$ and the LRT returns a significant p-value.

Production study in English: results



Production study in English: analyses

$$\log(P(\text{choosing singular})/P(\text{choosing plural})) = a + b \times n$$

where n is the number of boxes with a unique object.

Full model: $b = 0.34$
LRT confirms significant improvement in model fit compared to null model: $\chi^2(1) = 29.92$ and $p < 0.001$

Same analysis on the subset of **non-extreme conditions** (which represent *truly mixed* situations). **Model without extreme conditions:** $b = 0.24$
LRT still confirms significant improvement in model fit compared to null model: $\chi^2(1) = 9.20$ and $p = 0.002$

Core Methodological Question for comprehension studies

If gradient effects are also observed in comprehension...

Core Methodological Question

Experimentally, how can we disentangle readings from gradient effects?

Gradient effects are a **confounding factor** in determining which readings exist!

Why do gradient effects represent a confounding factor?

1. **Making it look like the weak reading exists when it does NOT:** the quantitative shift due to gradience can be mistaken for a qualitative shift due to an additional reading being made true.
2. Or conversely, **making it look like the weak reading does not exist when it DOES:** gradience can conceal a level of reading.

What is the source of gradient effects?
Maybe, proximity to the closest situation that satisfies a certain reading.

Outline

Some NPs experiment (continuous judgments): design and goals

We have conducted several experiments of language comprehension, asking for judgments (on a continuous scale) of a sentence relative to a picture.
The sentences from the all experiments followed the same structure :

Each box contains [plural expression].

Each experiment used a different plural expression:

- bare plurals
- *several NPs*
- *some NPs*

→ **I only present the *some NPs* experiment**, mainly because it is the only experiment to have one version with continuous judgments and one with binary judgments.
→ See Rong 2025 for an analysis of the remaining experiments.

Some NPs experiment (continuous judgments): design and goals

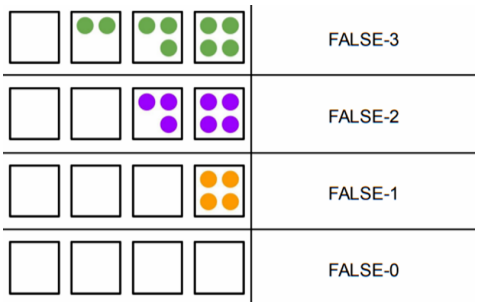
Main differences with the design of the production study:

1. Within-subjects instead of between-subjects.
2. Addition of literally false conditions (at least one box is empty).
3. 4 boxes instead of 10 boxes.
4. "Each" instead of "every" (to limit cumulative readings).
5. Geometric shapes instead of natural objects.

Main goal: observe whether there are **gradient effects within a same reading** (no such analysis in previous literature).

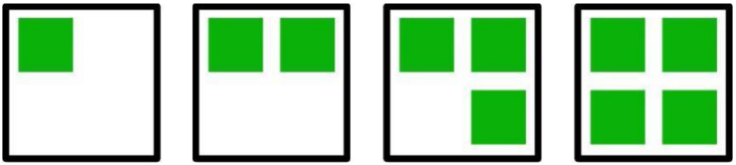
Some NPs experiment (continuous judgments): design and goals

A box containing several shapes is called a **strong verifier**.



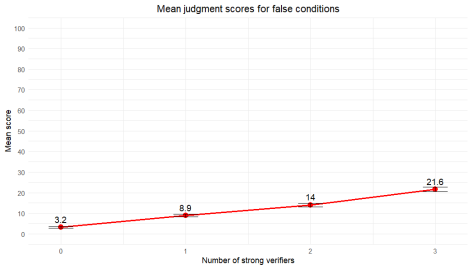
Some NPs experiment (continuous judgments): design and goals

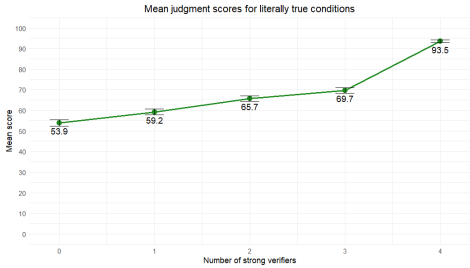
Example of a trial:



Some NPs experiment (continuous judgments): results

200 participants (after exclusions) recruited through Prolific.
Each participant saw each condition 3 times.

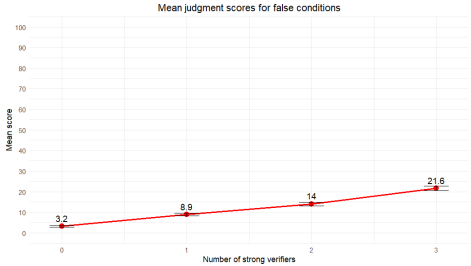


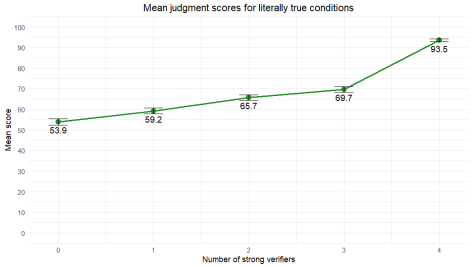


Some NPs experiment (continuous judgments): results

Visually:

- Qualitative shifts from **FALSE** to **LITERAL**, and from **WEAK** to **STRONG**.
- Quantitative shifts within **FALSE** and within **LITERAL**+**WEAK**.





Some NPs experiment (continuous judgments): analyses

Four predictors:

- c_{vrf} (number of strong verifiers)
- c_{lit} (binary variable indicating whether the condition is literally true)
- c_{weak} (indicating whether the condition supports a weak reading)
- c_{str} (indicating whether the condition supports a strong reading)

On literally true conditions: we fit a linear mixed-effects model predicting responses as a function of c_{vrf}, with random intercepts and slopes by participant.

$$\text{response} \sim c_{\text{vrf}} + (1 + c_{\text{vrf}} \mid \text{participant})$$

Result (as expected): a positive slope in the linear model and a significant LRT p-value (comparison with a null model). $\chi^2(1) = 1052.9$, $p < 10^{-15}$.
LRT on **WEAK** conditions alone: $\chi^2(1) = 65.19$, $p < 10^{-15}$.
$\rightarrow$ **Gradience is indeed present within the same reading.**

Some NPs experiment (continuous judgments): analyses

To identify the best-fitting combination of predictors, among the $2^4 = 16$ possible combinations: model comparisons across all 9 conditions using the Bayesian information criterion (BIC) and the Akaike information criterion (AIC).

Best-fitting model across all 9 conditions (according to both BIC and AIC):

$$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{str}} + (1 \mid \text{participant})$$

$\rightarrow$ Seems to favor approaches that **do not predict a weak reading.**

Second best-fitting model ($\Delta BIC = 9$):

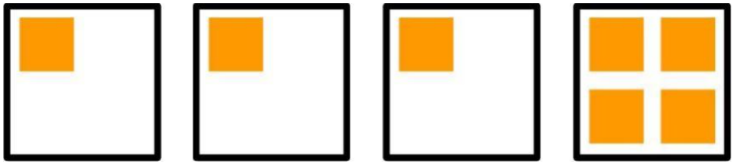
$$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$$

Outline

Some NPs experiment (binary judgments): design

Goal: are gradient effects still present with binary judgments?

Same sentences and pictures as in the version with continuous judgments.
Example of a trial:



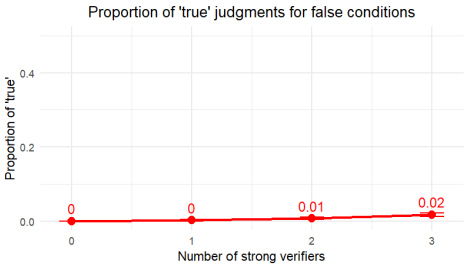
Do you think the sentence below is true or false?

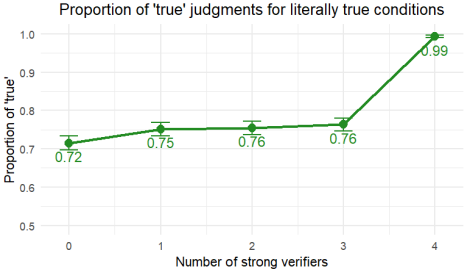
Each box contains some squares.

☐ false ☐ true

Some NPs experiment (binary judgments): results

200 participants (after exclusions) recruited through Prolific.
Each participant saw each condition 3 times.

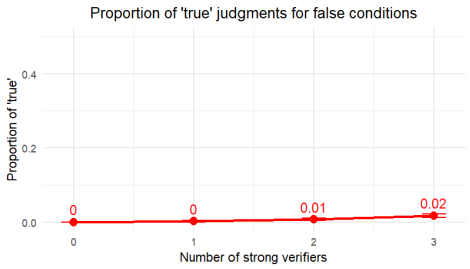


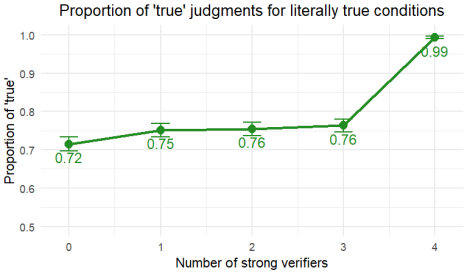


Some NPs experiment (binary judgments): results

Visually:

- Hardly any gradience within **FALSE** cases, or within **WEAK+LITERAL** cases.
- The only noticeable qualitative shifts: from **FALSE** to **LITERAL** and from **WEAK** to **STRONG**.





Some NPs experiment (binary judgments): analyses

Same four predictors:

- c_{vrf} (number of strong verifiers)
- c_{lit} (binary variable indicating whether the condition is literally true)
- c_{weak} (indicating whether the condition supports a weak reading)
- c_{str} (indicating whether the condition supports a strong reading)

Best-fitting model across all 9 conditions (according to both BIC and AIC):

$$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{str}} + (1 \mid \text{participant})$$

$\rightarrow$ Seems to favor approaches that **do not predict a weak reading.**

BUT this could well be due to **limitations of the logistic model**: exaggerated effects of tiny (insignificant) gradience within **FALSE**.

Some NPs experiment (binary judgments): analyses

Follow-up analysis: model comparison without c_{lit} on the subset of literally true conditions.

Best-fitting model across the subset of 5 conditions:

$$\text{response} \sim c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$$

This is the **only time** in all our English experiments that c_{weak} **was present in the best-fitting model.**

$\rightarrow$ **Could it be that continuous judgments generate more gradient effects which then conceal a level of reading?**

Outline

A language with optional number marking

A threefold number marking in Mandarin:

1. Bare noun: **number-neutral** (Zhang 2014; Cheng and Sybesma 1999...).

2. [one + CL] where CL is the '**singular**' classifier ≈ *a NP*
   In Mandarin: *yige* 一个、*yiben* 一本⋯⋯
   Triggers a **uniqueness** inference.

3. [one + *xie*] where *xie* 些 is the '**plural**' classifier ≈ *(some) NPs*
   In Mandarin: *yixie* 一些.
   Triggers a **multiplicity** inference.

It can be shown that all three forms have the **same truth conditions**.

Comprehension study in Mandarin

Picture stimuli identical to those of the English comprehension experiment, with translated instructions and the following stimulus sentence:

(4)

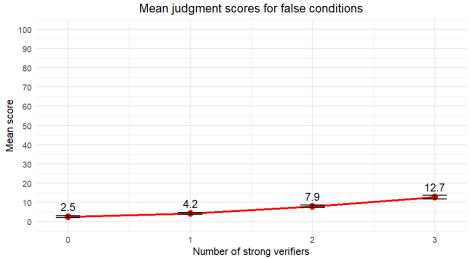
每	个	盒子	里	都	有	一	些	*[NP]*
měi	gè	hé-zi	lǐ	dōu	yǒu	yī	xiē	
each	CL	box	in	DOU	EXIST	**one**	**xie**	[NP]

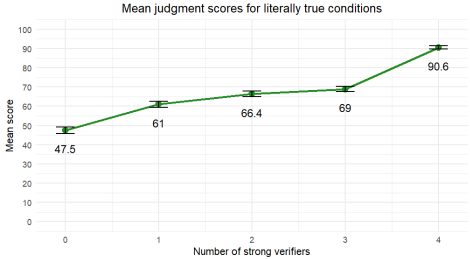
'Each box contains [one + *xie* + NP].'

xie experiment: results

Visually:

- Qualitative shifts from **FALSE** to **LITERAL**, from **WEAK** to **STRONG**, but also (it seems!) from **LITERAL** to **WEAK**.
- Quantitative shifts within **FALSE** and within **LITERAL**+**WEAK**.





xie experiment: analyses

Same four predictors:

- c_{vrf} (number of strong verifiers)
- c_{lit} (binary variable indicating whether the condition is literally true)
- c_{weak} (indicating whether the condition supports a weak reading)
- c_{str} (indicating whether the condition supports a strong reading)

On literally true conditions, we fitted a linear mixed-effects model:

$$\text{response} \sim c_{\text{vrf}} + (1 + c_{\text{vrf}} \mid \text{participant})$$

Result (as expected): a positive slope in the linear model and a significant LRT p-value (comparison with a null model). $\chi^2(1) = 858.13$, $p < 10^{-15}$.

LRT on **WEAK** conditions alone, with the same conclusions:
$\chi^2(1) = 38.34$, $p < 10^{-9}$.

xie experiment: analyses

Best-fitting model across all 9 conditions (according to both BIC and AIC):

$$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$$

Second best-fitting model ($\Delta BIC = 62$):

$$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{str}} + (1 \mid \text{participant})$$

$\rightarrow$ Contrary to the experiment on *some NPs* (continuous judgments), c_{weak} **is present in the best model.**

Outline

Answers to the Core Questions

Core Theoretical Question 1

What are the available readings?

- In every experiment, the **literal** and **strong** readings are present in the best model.
- At first glance, **with continuous judgments**, weak reading not supported in English comprehension (no improvement via c_{weak}).
- Best model: response $\sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{str}} + (1 \mid \text{participant})$
- Weak reading detected in binary *some NPs* task $\rightarrow$ methodological challenge.

HOI approach	{literal, weak, strong}
Zweig(+Ivlieva)'s approach	{strong} $\rightarrow$ not supported
Presuppositional Exhaustification approach	{literal(#), strong}
Homogeneity-based approach	{literal(#), strong}

Answers to the Core Questions

Core Theoretical Question 2

How universal are the mechanisms of plural interpretation?
More specifically, as a case study, what are the available readings in Mandarin, a language with optional number marking?

- Mandarin shows gradient effects *and* supports the weak reading (c_{weak} improves model fit).
- Further theoretical work needed on link between optional number marking systems and possible availability of weak reading.

Answers to the Core Questions

Core Methodological Question

Experimentally, how can we disentangle readings from gradient effects?

- Gradience modeled via strong verifier count (c_{vrf}) + binary factors for readings (c_{lit}, c_{weak}, c_{str}).
- Alternative: weights on readings + for each reading, distance to closest situation that satisfies the reading (e.g., Chemla and Spector 2014).

Remaining puzzles

- **Source of gradient effects:** verisimilitude vs. typicality (Chemla and Spector 2014, van Tiel and Geurts 2014).
- **Cumulative readings:** acceptability varies across plural expressions.
- **Binary vs. continuous responses:** modeling binary responses as a function of continuous responses.
- **Production vs. comprehension:** underexplored link.

Future directions

- Investigate **other lexical scales** with scalar implicatures.
- Extend to **other quantifiers**.
- Explore **context-sensitivity** by controlling the Question Under Discussion (QUD).
- Refine ambiguity resolution using **probabilistic models** (e.g., the Rational Speech Act framework).

Thank you!

References I

Ahn, Dorothy, Ankana Saha, and Uli Sauerland (2020). "Positively polar plurals: Theory and predictions". In: Semantics and linguistic theory, pp. 450–463.

Chemla, Emmanuel and Benjamin Spector (2014). "Distinguishing typicality and ambiguities, the case of scalar implicatures". Unpublished manuscript.

Cheng, Lisa Lai-Shen and Rint Sybesma (1999). "Bare and not-so-bare nouns and the structure of NP". In: Linguistic inquiry 30.4, pp. 509–542.

Enguehard, Emile (2024). "What number marking on indefinites means: conceivability presuppositions and sensitivity to probabilities". In: Sinn und Bedeutung. Vol. 28.

Ivlieva, Natalia (2020). "Dependent plurality and the theory of scalar implicatures: Remarks on Zweig 2009". In: Journal of Semantics 37.3, pp. 425–454.

References II

Jiang, Yizhen and Yasutada Sudo (2023). "Putting bare plurals into context". In: Hnm2—gaps and imprecision in natural language semantics: homogeneity effects

Križ, Manuel (2017). "Bare plurals, multiplicity, and homogeneity". Unpublished manuscript.

Rong, Claire (2025). "Plurals under quantification: new experimental perspectives". MA thesis. École Normale Supérieure.

Spector, Benjamin (2007). "Aspects of the Pragmatics of Plural Morphology: On Higher-Order Implicatures". In: Presupposition and Implicature in Compositional Semantics.

Stateva, Penka, Sara Andreetta, and Arthur Stepanov (2016). "On the nature of the plurality inference: Ladybugs for Anne". In: Papers dedicated to Anne Reboul. Lyon: CNRS.

References III

van Tiel, Bob and Bart Geurts (2014). "Truth and typicality in the interpretation of quantifiers". In: Sinn und Bedeutung. Vol. 18.

Zhang, Niina Ning (2014). "Expressing number productively in Mandarin Chinese". In: Linguistics.

Zweig, Eytan (2007). "Number-neutral bare plurals and the multiplicity implicature". In: Linguistics and Philosophy 32.