

École Normale Supérieure
Master in Cognitive Science
Major: Linguistics

**Plurals under quantification:
new experimental perspectives**

THESIS

Candidate: Claire RONG
Supervisor: Benjamin SPECTOR

June 2025

Declaration of originality

This thesis investigates the interpretation of plural indefinites under universal quantification. While the topic has been widely discussed in theoretical literature, this work introduces new experimental perspectives that help adjudicate between competing semantic theories.

The theoretical component of this thesis builds on established formal frameworks for plural interpretation, while the experimental component presents original data obtained through well-established methods in experimental pragmatics, including truth-value judgment tasks and sentence production. The experimental design is specifically tailored to detect gradient effects in truth-value judgments, a phenomenon observed but not accounted for in previous work (Chemla and Spector 2011; Stateva, Andreetta, and Stepanov 2016; Jiang and Sudo 2023). We conduct statistical model comparisons to find the best model predicting both gradient effects and the distribution of readings. The findings offer empirical support for certain semantic theories over others and open new directions for cross-linguistic investigation into number marking systems.

Declaration of contribution

The contributions are the following:

- definition of the research question: Benjamin Spector.
- literature review: Claire Rong, Benjamin Spector.
- choice of methodology: Benjamin Spector.
- designing the experiments: Claire Rong, Benjamin Spector.
- statistical analysis design: Benjamin Spector.
- implementation and coding of statistical analyses: Claire Rong (with the help of AI tools).
- interpretation of the results: Benjamin Spector, Claire Rong.
- writing the thesis: Claire Rong.
- proof-reading the thesis: Claire Rong, Benjamin Spector, Jakob Süskind, Léo Wang.

I have used AI tools (ChatGPT and DeepSeek) mainly for non-substantive tasks, including assistance with R programming. The only intellectual contribution from an AI tool (ChatGPT-o3) involved answering clarification questions related to advanced statistical analyses relevant to this thesis. These answers were either simply confirmatory or ultimately not implemented in the thesis.

Abstract

We investigate the interpretation of plural expressions, starting from the puzzle of the logical gap between their interpretations in affirmative and negative sentences. While bare plurals (e.g., “books”) typically receive an *at least two* reading in upward-entailing environments (e.g., “The box contains books”), they yield an *at least one* reading in downward-entailing contexts (e.g., “The box doesn’t contain books”). Existing bivalent and trivalent theoretical approaches differ in their predictions about truth conditions of plural indefinites under the scope of a universal quantifier (e.g., “Every box contains books”) in mixed scenarios (some boxes contain one book, others contain several).

Through experimental studies, we uncover gradient truth-value judgments in such mixed scenarios, which challenge the categorical predictions of all available theories. We propose a new model incorporating gradience along with factors coding for different readings of the quantified sentence. Additionally, we extend our investigation to Mandarin Chinese, a language with optional number marking, finding similar gradient effects but with different levels of accessed readings.

Our results raise key questions about (i) the empirical identification of accessible readings, (ii) methods to disentangle readings from gradient effects, and (iii) cross-linguistic variation in plural interpretation.

Acknowledgments

I would like to express my deepest thanks to my supervisor, Benjamin Spector. My academic trajectory and intellectual development owe more to Benjamin than I can express. He encouraged me to join the Cogmaster in the first place and has been the most supportive mentor one could imagine, in addition to being an academic role model and the person from whom I have learned the most linguistics. His patience, clarity, and intellectual generosity have been a driving force behind my decision to pursue linguistics during my master's and beyond. I hope to continue this path with the same rigor and integrity that Benjamin has always exemplified.

I wish to thank all my professors from the Cogmaster for training me not only in linguistics, but also more broadly in cognitive science. In particular, I have benefited from enlightening discussions and valuable support from Salvador Mascarenhas and Jeremy Kuhn over several projects throughout the past two years.

For filling my master's studies with friendship and wonderful memories, I thank the Linguistics Buddies: Baptiste L.-U., Baptiste S., Carlotta, Margot, Mattia, Nia, and Sarah. I am also most grateful to my friends from the Bateau, the Chalet, and the rest of Institut Jean Nicod for welcoming me into such a stimulating and fun environment.

Contents

Introduction	3
The puzzle of a logical gap	3
A tale of two readings	4
Illustration of the bivalent approaches	5
Illustration of the trivalent approaches	6
The critical case of mixed scenarios	7
Outline of the discussion	8
1 Theoretical approaches to plural interpretation	9
1.1 Implicature approach based on Higher-Order Implicatures	9
1.2 Implicature approach based on Zweig 2007 and Ivlieva 2020	14
1.3 Presuppositional Exhaustification approach	19
1.4 Homogeneity-based approach	20
1.5 Interim summary: predictions of the different theories	22
2 Experiments on plurals in the scope of universal quantifiers	24
2.1 Experimental studies in previous literature	24
2.1.1 Context-sensitivity of bare plurals	24
2.1.2 Gradient effects in production	25
2.2 Production study	27
2.2.1 Methods	27
2.2.2 Analyses	28
2.2.3 Results	29
2.2.4 Discussion	30
2.3 Comprehension study on bare plurals	30
2.3.1 Methods	31
2.3.2 Analyses	32
2.3.3 Results	34
2.3.4 Discussion	35
2.4 Comprehension study: cumulativity of different plural expressions	37
2.4.1 Methods	37
2.4.2 Analyses	37
2.4.3 Results	38

2.4.4	Discussion	39
2.5	Comprehension study on <i>several NPs</i>	39
2.5.1	Methods	40
2.5.2	Analyses	40
2.5.3	Results	41
2.5.4	Discussion	42
2.6	Comprehension study on <i>some NPs</i> : continuous judgments . . .	43
2.6.1	Methods	44
2.6.2	Analyses	44
2.6.3	Results	44
2.6.4	Discussion	46
2.7	Comprehension study on <i>some NPs</i> : binary judgments	47
2.7.1	Methods	47
2.7.2	Analyses	47
2.7.3	Results	48
2.7.4	Discussion	49
3	Comparison with Mandarin	53
3.1	A language with optional number marking	53
3.2	Comprehension study on <i>xie</i> : continuous judgments	54
3.2.1	Methods	54
3.2.2	Analyses	55
3.2.3	Results	56
3.2.4	Discussion	57
	Synthesis and closing remarks	59
	Summary and methodological discussions	59
	Conclusion	60
	Bibliography	63
A	AIC tables	66
B	Preregistrations	68

Introduction

The puzzle of a logical gap

The interpretation of plural expressions has been extensively studied from both theoretical and experimental perspectives. Among the motivations for studying how plural morphology contributes to meaning is a puzzle revealed by a logical gap between the meaning of English bare plurals in affirmative and negative sentences. Consider the following pair:

- (1) a. The box contains books.
- b. The box doesn't contain books.

In (1a), the noun “books” is typically interpreted to mean *at least two*. We will call this the *multiplicity* inference. However, in (1b), “books” does not give rise to a multiplicity inference and seems to receive an *at least one* reading: the sentence is true if and only if the box does not contain any books. If (1b) were the logical negation of (1a), (1b) would be true as soon as the box does not contain at least two books, in particular it must be true if the box contains exactly one book. This is not the case, as (1b) is generally judged false if the box contains exactly one book, which represents the logical gap in the interpretation of bare plurals.

More generally, the *at least one* reading arises in downward-entailing (DE) environments, i.e. environments that reverse the direction of logical entailment¹. Here are some examples of DE environments other than negation:

- (2) a. If the box contains books, it must be handled with care.
- b. No box contains books.
- c. The box has never contained books.

“Books” indeed receives an *at least one* reading in all of these sentences: (2a) suggests that as soon as the box contains at least one book, it must be handled with care; (2b) suggests that no box contains any books; (2c) suggests that the box has never contained even a single book. In contrast, “books” generally receives an *at*

¹Formally, f is downward-entailing if whenever $P \models Q$, then $f(Q) \models f(P)$.

least two reading in upward-entailing (UE) environments, i.e. environments that preserve the direction of logical entailment, such as simple affirmative sentences like (1a).

A tale of two readings

How can one account for the existence of two readings? Here are two possible hypotheses:

Hypothesis 1 Bare plurals are inherently ambiguous between the readings *at least one* and *at least two*.

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.

As pointed out in Spector 2007, Hypothesis 1 can be ruled out by examining bare plurals' interpretation in non-monotonic environments, i.e. environments that are neither UE nor DE, such as the scope of the quantifier “exactly *N*”. Consider the original sentence (3a). We replace “books” by the two possible unambiguous meanings, in (3b) and (3c). Spector shows that (3a) is not equivalent to either (3b) or (3c) and we summarize the arguments below.

- (3) a. Exactly one box contains books.
- b. Exactly one box contains at least one book.
- c. Exactly one box contains at least two books.

(3a) suggests that exactly one box contains at least one book, that the number of books in question is actually at least two, and that all other boxes do not contain books. If exactly one box contains two books and all other boxes contain one, (3a) is not intuitively judged true, but (3c) is. (3b) suggests that exactly one box contains any books at all and that all other boxes do not contain books. In particular, if exactly one box contains exactly one book and all other boxes do not contain books, (3b) is intuitively judged true, but both (3a) and (3c) are false. Thus, none of the sentences in (3) is equivalent to any other. If “books” were ambiguous between “at least one book” and “at least two books”, either (3b) or (3c) would have removed the ambiguity and captured the truth conditions of (3a), but this is not the case.

Let us take a closer look at Hypothesis 2, which is supported by several families of approaches to the semantic and pragmatic interpretation of plural indefinites, particularly bare plurals. Hypothesis 2 can be instantiated in two ways, both of which are found in the literature:

1. One class of approaches are bivalent approaches. They share the assumption that bare plurals have an *at least one* denotation, with the *at least two* reading arising via pragmatic strengthening. Using a bare plural in an UE environment to describe an *exactly one* situation is therefore logically true, but pragmatically weird.
2. Another class of approaches are trivalent approaches. They share the assumption that bare plurals are logically true in *at least two*-situations and logically false in *none*-situations. The logical gap of *at least one*-situations is captured by the ‘undefined’ truth-value.

The rest of the chapter gives an overview of the mechanisms from each class of approaches.

Illustration of the bivalent approaches

The mechanisms of bivalent approaches are all implicature-based, as they make explicit use of scalar implicatures (Sauerland 2003; Spector 2007; Zweig 2007). The scalar implicatures involved are based on the $\langle \text{PL}, \text{SG} \rangle$ scale, although they differ by their other assumptions and their implementation of the pragmatic competition. More precisely, the $\langle \text{PL}, \text{SG} \rangle$ scale should be written as $\langle \text{NP}_{\text{PL}}, \text{a NP}_{\text{SG}} \rangle$ in the case indefinites in English. We illustrate the reasoning by drawing a parallel with a well-known example of scalar implicature, $\langle \text{some}, \text{all} \rangle$.

- (4) a. I read some of the books.
 b. I read all of the books.

In its literal sense, (4b) asymmetrically entails (4a). A literal listener who hears (4a) will only infer that the speaker read at least one book. For a pragmatic listener, however, (4a) stands in competition with a sentence that the speaker could have said but did not say, namely (4b). Following Grice’s Maxim of Quantity², if the speaker is truthful and if they actually read all the books, then they would have used (4b). The listener assumes that the speaker is truthful and is not in a position to assert (4b), deducing that it must not be the case that the speaker read all the books.

The same reasoning applies for the scale $\langle \text{NP}_{\text{PL}}, \text{a NP}_{\text{SG}} \rangle$:

- (5) a. I read books.
 b. I read a book.

²Maxim of Quantity: “Make your contribution as informative as is required (for the current purposes of the exchange). Do not make your contribution more informative than is required.”

We assume that “books” literally means *at least one* book. Let us also make the (unrealistic) assumption that “a book” receives the interpretation of “exactly one book” (I will call this the *uniqueness* inference).³ Under this interpretation, (5b) asymmetrically entails (5a). Because the bare plural is in competition with (i.e. is on the same scale as) a singular noun, a pragmatic listener infers that the speaker is not in a position to assert (5b). Then, the meaning of (5a) gets strengthened to “I read at least two books”, hence the multiplicity inference. Across all implicature-based accounts, there is consensus that the multiplicity inference triggered by plural indefinites is not a standard entailment, as the plural meaning is not part of the content that is negated in simple negative sentences like (1b).

Illustration of the trivalent approaches

The second class of approaches includes the homogeneity-based account (Križ 2017) and the presuppositional exhaustification account (Bassi, Del Pinal, and Sauerland 2021; Ahn, Saha, and Sauerland 2020). The latter incorporates elements from the theory of scalar implicature, but ultimately adopts a trivalent semantics and makes the same predictions as the homogeneity-based approach. This paragraph will focus on presenting the homogeneity-based proposal, as it has more intuitive illustrations than presuppositional exhaustification. According to the homogeneity-based proposal, bare plurals have a strictly plural denotation, and *at least one* readings arise as a result of contextual factors. For example:

Context A: I have an assignment asking me to read at least one book, I did the bare minimum and read exactly one. I claim:

(6) I read books.

Context B: I am not allowed to read any books from the Restricted Section of the Hogwarts Library, but I snuck in and read exactly one book. I claim:

(7) I didn’t read books.

In context A, (6) is undefined in the trivalent framework posited by the homogeneity-based theory. However, (6) can be judged “true enough for current purposes” (to use the authors’ expression from Križ and Spector 2021), as the fact that (6) is *not true* is not relevant for the Question Under Discussion (QUD). Similarly, in context B, (7) is undefined but can be judged false, because of the QUD is whether

³Clearly, this interpretation cannot be a plausible semantics for *a NP*. This can be shown by embedding the singular indefinite in a DE environment. For instance, if “a book” meant “exactly one book”, the sentence “If you read a book, you will gain knowledge” implies that you do not gain knowledge if you have read more than one book, which is not a desired prediction. The next chapter will discuss how the *exactly one* interpretation is derived through implicature.

I have read any books at all. These two contexts show how utterances that are logically undefined can be conflated sometimes with true cases and sometimes with false cases.

The critical case of mixed scenarios

All the accounts we have mentioned differ, sometimes even within the same class of approaches, with respect to the truth conditions of a plural indefinite under the scope of a universal quantifier, as in sentence (8). Crucially, they predict different truth-values for (8) in *mixed* scenarios, where some boxes contain exactly one book and others multiple. Figure 1 gives an example of a mixed scenario.

(8) Every box contains books.



Figure 1: Example of a mixed scenario for “every box contains books”

Imagine now that we are not asked to provide a categorical truth-value judgment aligning with one theory or another. Instead, we are asked to indicate our judgment of how well the quantified sentence describes the scenario, using a cursor on a continuous scale. Intuitively, one would place the cursor somewhere in the middle, rather than at either extreme. Moreover, the cursor’s placement would likely vary if the scenario presented a different distribution while still being mixed: for example, if one box contains a single book and nine boxes contain several, or vice versa. The phenomenon where continuous truth-value judgments vary with the distribution within a mixed scenario will be called a *gradient* effect.

In line with our intuitions, our experimental results (Chapter 2) reveal gradient effects in truth-value judgments of mixed scenarios, when participants provide judgments on a continuous scale. We may wonder whether this is due to the response option being continuous. Interestingly, gradient effects are also observed in production and in comprehension tasks with binary judgments. At any rate, gradient effects are not predicted by existing theories.

Even without taking into account gradient effects, a theoretical question arises because theories differ in the readings they predict, that is, the truth-conditional interpretations they assign to sentences like (8). This leads to our first core theoretical question, which our experimental data will help answering:

Core Theoretical Question 1

What are the available readings?

However, in practice, gradient effects pose a challenge for identifying these readings: the continuous nature of truth-value judgments can blur the boundary between different readings. When a new reading becomes available, its effect on judgments may be masked by the quantitative variation already induced by gradience. This leads us to our core methodological question:

Core Methodological Question

Experimentally, how can we disentangle readings from gradient effects?

This thesis also explores cross-linguistic variation in plural interpretation. Do gradient effects and plural readings differ between languages with obligatory number marking (such as English) and those without? We investigate the example of Mandarin Chinese, a language where number marking is optional and where plurality is expressed via classifiers and suffixes.

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 of the discussion

This thesis is organized as follows. In Chapter 1, we present in detail the main theoretical approaches to plural interpretation and we show that they make different truth-value predictions for both non-quantified and universally quantified sentences. Chapter 2 reports new experimental findings in language production and comprehension. Observations that are not predicted by existing theories include gradient effects in truth-value judgments for several types of indefinite plural expressions, not only across readings, but most importantly, within readings. Chapter 3 provides an empirical comparison with Mandarin, a language with optional number marking and featuring number-neutral morphology. The conclusion chapter includes methodological discussions and suggests ideas for future research.

Chapter 1

Theoretical approaches to plural interpretation

1.1 Implicature approach based on Higher-Order Implicatures

Based on Spector 2007, we present a reformulation of the Higher-Order Implicatures (HOI) approach using an exhaustivity operator **EXH**. **EXH** is an operator which takes as input an utterance and returns its strengthened meaning. In what follows and contrary to the original paper, **EXH** is not part of the metalanguage but is instead encoded directly in the syntax.

For a sentence p , its first-order exhaustified (i.e. strengthened) meaning **EXH**(p) is given by (9)¹:

$$(9) \quad \mathbf{EXH}(p) \equiv \text{LIT}(p) \wedge \bigwedge_{\substack{q \in \text{ALT}^*(p) \\ p \neq q}} \neg q$$

- $\text{LIT}(p)$ denotes the proposition expressing its literal meaning;
- $\text{ALT}(p)$ denotes the set of p 's scalar alternatives;
- $\text{ALT}^*(p)$ denotes the set of the propositions expressed by the elements of $\text{ALT}(p)$.

Taking the example of the sentence “I read some of the books” and the lexical scale $\langle \text{some}, \text{all} \rangle$, we have:

$$(10) \quad \begin{aligned} \text{LIT}(\text{I read some of the books}) &= \text{I read at least one book} \\ \text{ALT}(\text{I read some of the books}) &= \{\text{I read all of the books}\} \\ \text{ALT}^*(\text{I read all of the book}) &= \{\text{I read all of the books}\} \end{aligned}$$

¹Throughout this whole chapter, we are giving schematic derivations and conflating use and mention in our formulas. Strictly speaking, we should use double brackets $\llbracket \rrbracket$ in our formulas.

“I read at least one book” does not entail “I read all of the book”, therefore:
 $\mathbf{EXH}(\text{I read some of the books}) \equiv (\text{I read at least one book}) \wedge \neg(\text{I read all of the books})$
 $\equiv \text{I read some but not all of the books}$

In (10), one application of \mathbf{EXH} is sufficient to derive the intuitively correct strengthened meaning. However, in the case of the $\langle \text{NP}_{\text{PL}}, \text{a NP}_{\text{SG}} \rangle$ scale, we will see that one application of \mathbf{EXH} is not enough. Consider (5), repeated here:

- (11) a. I read books.
 b. I read a book.

The implementation is based on a mereological theory of plurality (following Link et al. 1983; Link 1990; Landman 2000, a.o.) where plural nominal morphology is a function that takes as input a set P of atomic individuals and returns its closure under sum, notated $\oplus P$. For example:

- (12) $\llbracket \text{book} \rrbracket = \{a, b, c\}$
 $\llbracket \text{books} \rrbracket = \llbracket \text{book-PL} \rrbracket = \oplus \llbracket \text{book} \rrbracket = \{a, b, c, a \oplus b, b \oplus c, a \oplus c, a \oplus b \oplus c\}$

In this sense, a bare plural is semantically number-neutral and has the literal meaning *at least one*.

To derive the multiplicity inference from (11a), the informal reasoning is as follows. The literal meaning of (11a) is “I read at least one book”. In order for its meaning to get strengthened to “I read more than one book”, (11b) needs to have a competitor whose meaning is “I read exactly one book”, in order to negate that competitor. “I read exactly one book” is neither a plausible alternative (for reasons of syntactic complexity) to (11a), nor the literal meaning of (11a)’s actual alternative, (11b). However, “I read exactly one book” is the *strengthened* meaning of (11b). Thus, (11a) is strengthened relative to the strengthened meaning of (11b). Because two strengthenings take place, the pragmatic meaning is said to be derived through *higher-order* implicature. Syntactically, two strengthenings are implemented by two iterations of \mathbf{EXH} .

The formal derivation is as follows. If \mathbf{EXH} is applied once to (11a), no strengthening takes place:

- (13) $\text{LIT}(\text{I read books}) = \text{I read at least one book}$
 $\text{ALT}(\text{I read books}) = \{\text{I read a book}\}$
 $\text{ALT}^*(\text{I read books}) = \{\text{I read at least one book}\}$
 The only element in $\text{ALT}^*(\text{I read books})$ is the literal meaning of –and therefore entailed by– “I read books”. Thus, $\mathbf{EXH}(\text{I read books}) \equiv \text{LIT}(\text{I read books})$.

However, the meaning gets strengthened after a second application of \mathbf{EXH} :

$$(14) \quad \mathbf{EXH}(\mathbf{EXH}(p)) \equiv \text{LIT}(\mathbf{EXH}(p)) \wedge \bigwedge_{\substack{q \in \text{ALT}^*(\mathbf{EXH}(p)) \\ \mathbf{EXH}(p) \not\models q}} \neg q$$

In (14), note that:

1. $\mathbf{EXH}(p)$ is equivalent to its literal meaning, because it is already exhausted. We are writing $\text{LIT}(\mathbf{EXH}(p))$ to keep the parallel with (9).
2. As \mathbf{EXH} is implemented in the syntax, we have

$$\text{ALT}(\mathbf{EXH}(p)) = \{\mathbf{EXH}(q), q \in \text{ALT}(p)\}.$$

Therefore, the big conjunction is equivalent to

$$\bigwedge_{\substack{q \in \text{ALT}(p) \\ \mathbf{EXH}(p) \not\models \mathbf{EXH}(q)}} \neg \mathbf{EXH}(q)$$

We will be using this second expression in our derivations below, as the computations are slightly simpler.

We first compute the (first-order) strengthened meaning of every element in $\text{ALT}(\text{I read books})$, i.e. of its only element “I read a book”.

$$\begin{aligned}
(15) \quad & \text{LIT}(\text{I read a book}) = \text{I read at least one book} \\
& \text{ALT}(\text{I read a book}) = \{\text{I read several books}\}^2 \\
& \text{This is assuming that } \langle \text{a, several} \rangle \text{ also forms a lexical scale.} \\
& \text{ALT}^*(\text{I read a book}) = \{\text{I read several books}\} \\
& \text{“I read at least one book” does not entail “I read several books”, therefore:} \\
& \mathbf{EXH}(\text{I read a book}) \equiv (\text{I read at least one book}) \wedge \neg(\text{I read several books}) \\
& \equiv \text{I read exactly one book}
\end{aligned}$$

Then, using (14), we compute the second-order strengthened meaning of (11a). Given that $\mathbf{EXH}(11a)$ does not entail $\mathbf{EXH}(\text{I read a book})$, we can negate $\mathbf{EXH}(\text{I read a book})$:

$$\begin{aligned}
(16) \quad & \mathbf{EXH}(\mathbf{EXH}(\text{I read books})) \equiv \text{LIT}(\mathbf{EXH}(\text{I read books})) \wedge \neg \mathbf{EXH}(\text{I read a book}) \\
& \equiv \text{LIT}(\text{I read at least one book}) \wedge \neg(\text{I read exactly one book}) \\
& \equiv \text{I read several books}
\end{aligned}$$

Having illustrated the mechanism on scalar implicatures in non-quantified sentences, let us now turn to quantified sentences. For the kind of quantified

²Here and in the rest of the paper, we will use “several” with the meaning *at least two*, but we acknowledge the ordinary usage of “several” to rather express a number greater or equal to 3. For instance, the *Cambridge Dictionary of English* offers the definition: “more than two and fewer than many”.

sentences p that we are dealing with, it will only be necessary to compute first-order ($\mathbf{EXH}(p)$) and second-order ($\mathbf{EXH}(\mathbf{EXH}(p))$) iterations. An explanation will also be provided as to why it is not necessary to compute higher orders for the case at hand. In the formulas below, we will write “every...BP” (BP for bare plural) and “every...a” as shorthand for “every box contains books” and “every box contains a book”.

If \mathbf{EXH} is applied once to “every...BP” and if one only considers the scale $\langle \text{BP}, \text{a NP} \rangle$, no strengthening takes place:

- (17) $\text{LIT}(\text{every...BP}) = \text{every...one or more}$
 $\text{ALT}(\text{every...BP}) = \{\text{every...a}\}$
 $\text{ALT}^*(\text{every...BP}) = \{\text{every...one or more}\}$
The only element in $\text{ALT}^*(\text{every...BP})$ is the literal meaning of –and therefore entailed by– “every...BP”. Thus, $\mathbf{EXH}(\text{every...BP}) \equiv \text{LIT}(\text{every...BP})$.

This result is unchanged if we take into account both scales $\langle \text{BP}, \text{a NP} \rangle$ and $\langle \text{every}, \text{some} \rangle$:

- (18) $\text{ALT}(\text{every...BP}) = \{\text{every...a}, \text{some...BP}, \text{some...a}\}$
 $\text{ALT}^*(\text{every...BP}) = \{\text{every...one or more}, \text{some...one or more}\}$
All the elements in $\text{ALT}^*(\text{every...BP})$ are already entailed by “every...BP” and the meaning does not get strengthened.

Contrary to (17) and (18), two applications of \mathbf{EXH} will derive different readings depending on the scales under consideration.

Firstly, with $\langle \text{BP}, \text{a} \rangle$ only, we have $\text{ALT}(\text{every...BP}) = \{\text{every...a}\}$. We first compute the exhaustified meaning of “every...a”.

- (19) $\text{ALT}(\text{every...a}) = \{\text{every...several}\}$
 $\text{ALT}^*(\text{every...a}) = \{\text{every...several}\}$
“Every...several” is not entailed by “every...a”. Thus,
 $\mathbf{EXH}(\text{every...a}) \equiv (\text{every...one or more}) \wedge \neg(\text{every...several})$
 $\equiv \text{every...exactly one}$

As $\mathbf{EXH}(\text{every...a})$ is not entailed by “every...BP”, we have

- (20) $\mathbf{EXH}(\mathbf{EXH}(\text{every...BP})) \equiv (\text{every...one or more})$
 $\wedge \neg((\text{every...one or more}) \wedge \neg(\text{every...several}))$
 $\equiv \text{every...several}$

Secondly, if we take into account both scales $\langle \text{BP}, \text{a NP} \rangle$ and $\langle \text{every}, \text{some} \rangle$, then

$$\text{ALT}(\text{every...BP}) = \{\text{every...a}, \text{some...BP}, \text{some...a}\}$$

As above, we compute the exhaustified meaning of each element of $\text{ALT}(\text{every...BP})$

- (21) a. $ALT(\text{every...}a) = \{\text{every...BP, every... several, some...BP, some...}a, \text{some...several}\}$
 $ALT^*(\text{every...}a) = \{\text{every...one or more, every... several, some...one or more, some...several}\}$
The alternatives that are not entailed by “every...a” are “every...several” and “some...several”. Thus,
 $EXH(\text{every...}a) \equiv (\text{every...one or more}) \wedge \neg(\text{every...several}) \wedge \neg(\text{some...several})$
 $\equiv \text{every...exactly one}$
- b. $ALT(\text{some...BP}) = \{\text{some...}a, \text{every...BP, every...}a\}$
 $ALT^*(\text{some...BP}) = \{\text{some...one or more, every...one or more}\}$
The alternative that is not entailed by “some...BP” is “every...one or more”. Thus,
 $EXH(\text{some...BP}) \equiv (\text{some...one or more}) \wedge \neg(\text{every...one or more})$
 $\equiv \text{some but not all...one or more}$
- c. $ALT(\text{some...}a) = \{\text{some...NP, some...several, every...BP, every...}a, \text{every...several}\}$
 $ALT^*(\text{some...}a) = \{\text{some...one or more, some...several, every...one or more, every...several}\}$
The alternatives that are not entailed by “some...a” are “some...several”, “every...one or more” and “every...several”. Thus,
 $EXH(\text{some...}a) \equiv (\text{some...one or more}) \wedge \neg(\text{some...several}) \wedge \neg(\text{every...one or more})$
 $\wedge \neg(\text{every...several})$
 $\equiv (\text{some...exactly one}) \wedge \neg(\text{every...one or more})$

We have

$$ALT(EXH(\text{every...BP})) = \{\text{every...exactly one, some but not all...one or more, (some...exactly one)} \wedge \neg(\text{every...one or more})\}$$

None of the elements in $ALT(EXH(\text{every...BP}))$ is entailed by “every...BP”, therefore

$$(22) \quad \begin{aligned} EXH(EXH(\text{every...BP})) &\equiv (\text{every...one or more}) \\ &\quad \wedge \neg(\text{every...exactly one}) \\ &\quad \wedge \neg(\text{some but not all...one or more}) \\ &\quad \wedge \neg((\text{some...exactly one}) \wedge \neg(\text{every...one or more})) \\ &\equiv (\text{every...one or more}) \wedge \neg(\text{every...exactly one}) \end{aligned}$$

We will call this second reading the *weak* reading, as opposed to the reading “every...several” which will be called the *strong* reading. To summarize:

Scales used to generate ALT	EXH(every...BP)	EXH(EXH(every...BP))
$\langle \text{BP}, a \text{ NP} \rangle$	every...one or more	Strong reading: every... several
$\langle \text{BP}, a \text{ NP} \rangle$ and $\langle \text{some}, \text{every} \rangle$		Weak reading: (every...one or more) $\wedge \neg$ (every...exactly one)

Some final remarks about the HOI approach:

1. “To be a scale-mate of” is not assumed to be a transitive relation. As we have seen, *a NP* can be considered a scale-mate of a bare plural, and of *several NPs*, separately. However, bare plurals are not taken to be scale-mates of *several NPs*.
2. This approach posits ambiguity in the choice of possible lexical scales, ultimately resulting in ambiguity between possible readings.
3. Spector 2007 defines, for any sentence p , its closure under iterated alternatives³ $\overline{\text{ALT}}(p)$:

$$\overline{\text{ALT}}(p) = \bigcup_{n=0}^{+\infty} \text{ALT}^n(p)$$

where $\text{ALT}^0(p) = \text{ALT}(p)$ and $\text{ALT}^{n+1}(p) = \bigcup_{q \in \text{ALT}^n(p)} \text{ALT}(q)$.

It is shown in Spector 2007 that $\text{EXH}^{\text{on}}(p)$ ⁴ always stabilizes after a certain rank if $\overline{\text{ALT}}(p)$ is finite. While we are not concerned with reproducing this result using the syntactic operator **EXH**, it is sufficient for our purpose here to point out that $\overline{\text{ALT}}(\text{every...BP})$ is clearly finite, given that at most three scales ($\langle \text{BP}, a \text{ NP} \rangle$, $\langle \text{some}, \text{every} \rangle$ $\langle a, \text{several} \rangle$) are used to generate the successive $\text{ALT}^n(\text{every...BP})$. In particular, the sequence $\text{EXH}^{\text{on}}(\text{every...BP})$ stabilizes after index 2, which means that no more readings can be derived other than the weak and strong readings described above.

1.2 Implicature approach based on Zweig 2007 and Ivlieva 2020

We will review Zweig’s system described in Zweig 2007 and Zweig 2008, with additions from Ivlieva 2020 related to event summation. Also assuming the number-neutral denotation of bare plurals, the initial proposal was aimed at

³The name given in the original paper is *transitive closure*.

⁴ on denotes the n^{th} iterated application of **EXH**.

providing an explanation for the interpretation for the dependent plural reading of sentences such as “Three boys saw dogs”. Dependent plurality captures the requirement that there must be a plurality of dogs that were seen in total. In particular, the sentence is not true if all boys saw the same dog, even though it is the case for each boy that he saw “dogs”, with the number-neutral denotation of “dogs”. Zweig’s proposal uses neo-Davidsonian event semantics to derive the dependent plurality requirement. For simplicity, we will be using Davidsonian event semantics, as it is sufficient to introduce the system’s predictions. The kind of universally quantified sentences relevant for our interest (“Every box contains books”) does not involve dependent plurals, but it is nevertheless useful to examine the predictions of Zweig’s implicature system for our sentences.

The key assumptions of the system are:

1. For the choice of sites of implicature calculation, Zweig posits that alternatives can be generated at every scope site.
 2. For the generation of alternatives, it is assumed that ${}^{\oplus}P(X)$ has the alternative ${}^{\oplus}P(X) \wedge \text{ATOM}(X)$.
 3. For the exhaustification criteria, Zweig uses the following principle from Chierchia 2006, p.548:
- (23) In enriching a meaning, accord preference to the strongest option (if there is nothing in the context/common ground that prevents doing so).

We will go over these points one by one, to show how it translates formally for “Every box contains books”.

Firstly, there are three possible sites of implicature calculation. The denotation of our sentence before event closure is:

$$(24) \quad \lambda e'. e' \in \sum_{x \in \text{box}} \lambda e. \exists B [{}^{\oplus}\text{book}(B) \wedge \text{contain}(e, x, B)]$$

Starting at the lowest level, implicature calculations can take place at $\lambda e, \lambda e'$, and after event closure.

Secondly, regarding the alternatives generated, (25a) is assumed to have the alternative (25b):

- (25) a. $\lambda e. \exists B [{}^{\oplus}\text{book}(B) \wedge \text{contain}(e, x, B)]$
b. $\lambda e. \exists B [{}^{\oplus}\text{book}(B) \wedge \text{ATOM}(B) \wedge \text{contain}(e, x, B)]$

Note that Zweig considers alternatives in logical form, and not necessarily in meaning. Likewise, (26a) is assumed to have the alternative (26b):

$$\begin{aligned}
(26) \quad a. \quad & \lambda e'. e' \in \sum_{x \in \text{box}} \lambda e. \exists B [\oplus \text{book}(B) \wedge \text{contain}(e, x, B)] \\
b. \quad & \lambda e'. e' \in \sum_{x \in \text{box}} \lambda e. \exists B [\oplus \text{book}(B) \wedge \text{ATOM}(B) \wedge \text{contain}(e, x, B)]
\end{aligned}$$

Thirdly, for the exhaustification criteria, it is necessary to define a version of **EXH** that can apply at the predicate level, and not only at the sentence level as in the HOI approach. We adapt Ivlieva 2020’s reformulation of **EXH**⁵:

$$(27) \quad \llbracket \mathbf{EXH} \rrbracket \equiv \lambda P_{\langle \alpha, t \rangle}. \lambda x_{\alpha}. P(x) \wedge \bigwedge_{\substack{Q \in \text{ALT}^*(P) \\ Q(x) \models P(x)}} \neg Q(x)$$

We repeat Ivlieva’s reformulation of principle (23)⁶:

(29) **The Strongest Candidate Principle**

In choosing the correct meaning of a sentence with scalar items between candidates which differ with respect to where **EXH** is inserted, pick the strongest one, if that is possible.

As was said above, there are three scope sites in “Every box contains books” where **EXH** can be inserted. This results in a set of at most $2^3 = 8$ meanings, among which the strongest one is chosen, following the Strongest Candidate Principle.

At the lowest scope site, assuming the alternatives in (25), we can negate (25b) as it is strictly stronger than (25a).

$$\begin{aligned}
(30) \quad & \mathbf{EXH}(\lambda e. \exists B [\oplus \text{book}(B) \wedge \text{contain}(e, x, B)]) \\
& \equiv \lambda e. \exists B [\oplus \text{book}(B) \wedge \text{contain}(e, x, B)] \wedge \neg \exists B [\oplus \text{book}(B) \wedge \text{ATOM}(B) \wedge \\
& \quad \text{contain}(e, x, B)] \\
& \equiv \lambda e. \exists B [\oplus \text{book}(B) \wedge |B| > 1 \wedge \text{contain}(e, x, B)]
\end{aligned}$$

⁵Note that Ivlieva includes in the enriched meaning the negation of all *stronger* alternatives and not just *non-weaker* ones, as was the case in Spector’s approach. We keep these two different versions as given by their authors, but for our case, nothing crucial hinges on deciding between stronger or non-weaker alternatives.

⁶In Ivlieva 2020, the author amends this principle and proposes a Non-Weakening Condition in order to resolve issues related to dependent plurals. As this will not alter our conclusions for the case at hand, we will stick to the original Strongest Candidate Principle. We cite the amended principle for reference:

(28) **The Non-Weakening Condition (NWC)**

Do not introduce **EXH** in a structure S , if it would lead to a sentence meaning that is equivalent to or weaker than the meaning of S without that **EXH**.

At the next scope site, **EXH** may or may not have been applied at the previous point, which yields the following possible meanings:

$$\begin{aligned}
 (31) \quad a. \quad & \mathbf{EXH}(\lambda e'. e' \in \sum_{x \in \text{box}} \lambda e. \exists B [\oplus \text{book}(B) \wedge \text{contain}(e, x, B)])^7 \\
 & \equiv \lambda e'. e' \in \sum_{x \in \text{box}} \lambda e. \exists B [\oplus \text{book}(B) \wedge \text{contain}(e, x, B)] \\
 & \wedge \neg e' \in \sum_{x \in \text{box}} \lambda e. \exists B [\oplus \text{book}(B) \wedge \text{ATOM}(B) \wedge \text{contain}(e, x, B)]
 \end{aligned}$$

After event closure, this is equivalent to the weak reading: every box contains one or more books and it is not the case that every box contains exactly one book.

$$\begin{aligned}
 b. \quad & \mathbf{EXH}(\lambda e'. e' \in \sum_{x \in \text{box}} \mathbf{EXH}(\lambda e. \exists B [\oplus \text{book}(B) \wedge \text{contain}(e, x, B)])) \\
 & \equiv \lambda e'. e' \in \sum_{x \in \text{box}} \mathbf{EXH}(\lambda e. \exists B [\oplus \text{book}(B) \wedge \text{contain}(e, x, B)]) \\
 & \wedge \neg e' \in \sum_{x \in \text{box}} \mathbf{EXH}(\lambda e. \exists B [\oplus \text{book}(B) \wedge \text{ATOM}(B) \wedge \text{contain}(e, x, B)])
 \end{aligned}$$

Note that the first **EXH** returns $\lambda e. \exists B [\oplus \text{book}(B) \wedge |B| > 1 \wedge \text{contain}(e, x, B)]$, which is not entailed by what the second **EXH** returns, namely $\lambda e. \exists B [\oplus \text{book}(B) \wedge \text{ATOM}(B) \wedge \text{contain}(e, x, B)]$. Therefore, exhaustification at this step is void.

If **EXH** is applied after event closure, the set of possible meanings is:

$$\begin{aligned}
 (32) \quad a. \quad & \mathbf{EXH}(\exists e'. e' \in \sum_{x \in \text{box}} \mathbf{EXH}(\lambda e. \exists B [\oplus \text{book}(B) \wedge \text{contain}(e, x, B)])) \\
 & \equiv \mathbf{EXH}(\exists e'. e' \in \sum_{x \in \text{box}} \lambda e. \exists B [\oplus \text{book}(B) \wedge |B| > 1 \wedge \text{contain}(e, x, B)])
 \end{aligned}$$

Again, exhaustification at this step is void for the same reasons as in (31b).

$$\begin{aligned}
 b. \quad & \mathbf{EXH}(\exists e'. e' \in \sum_{x \in \text{box}} \lambda e. \exists B [\oplus \text{book}(B) \wedge \text{contain}(e, x, B)]) \\
 & \equiv \exists e'. e' \in \sum_{x \in \text{box}} \lambda e. \exists B [\oplus \text{book}(B) \wedge \text{contain}(e, x, B)] \\
 & \wedge \neg \exists e'. e' \in \sum_{x \in \text{box}} \lambda e. \exists B [\oplus \text{book}(B) \wedge \text{ATOM}(B) \wedge \text{contain}(e, x, B)]
 \end{aligned}$$

This corresponds to the weak reading.⁸

⁷Definition of an event sum: if P_1 and P_2 are two predicates of events, $P_1 + P_2$ is the predicate true of an event of the form $e_1 \oplus e_2$, where P_1 is true of e_1 and P_2 is true of e_2 .

⁸Compare this expression with (31b) after event closure, i.e.
 $\exists e'. e' \in \sum_{x \in \text{box}} \lambda e. \exists B [\oplus \text{book}(B) \wedge \text{contain}(e, x, B)] \wedge \neg e' \in \sum_{x \in \text{box}} \lambda e. \exists B [\oplus \text{book}(B) \wedge \text{ATOM}(B) \wedge \text{contain}(e, x, B)]$

To summarize, Table 1.1 gives the set of candidate meanings considering all three possible scope sites. We can see that the candidate meanings generated from 8 logical forms fall into 3 classes of equivalence, corresponding to the 3 readings already encountered in the HOI approach:

1. **Literal** reading: every box contains one or more books

$$\text{Logical form: } \exists e'. e' \in \sum_{x \in \text{box}} \lambda e. \exists B [\oplus \text{book}(B) \wedge \text{contain}(e, x, B)]$$

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

Two logical forms, equivalent in this case:

$$\exists e'. e' \in \sum_{x \in \text{box}} \lambda e. \exists B [\oplus \text{book}(B) \wedge \text{contain}(e, x, B)] \wedge \neg \exists e'. e' \in \sum_{x \in \text{box}} \lambda e. \exists B [\oplus \text{book}(B) \wedge \text{ATOM}(B) \wedge \text{contain}(e, x, B)]$$

and

$$\exists e'. e' \in \sum_{x \in \text{box}} \lambda e. \exists B [\oplus \text{book}(B) \wedge \text{contain}(e, x, B)] \wedge \neg e' \in \sum_{x \in \text{box}} \lambda e. \exists B [\oplus \text{book}(B) \wedge \text{ATOM}(B) \wedge \text{contain}(e, x, B)]$$

3. **Strong** reading: every box contains several books.

$$\text{Logical form: } \exists e'. e' \in \sum_{x \in \text{box}} \lambda e. \exists B [\oplus \text{book}(B) \wedge |B| > 1 \wedge \text{contain}(e, x, B)]$$

The Strongest Candidate Principle leads to the conclusion that the strong reading is the only one predicted in Zweig's approach. Note that there are four possible logical forms that generate this reading: as long as **EXH** is applied at the lowest scope site, it does not matter whether there are additional insertions of **EXH** in higher sites. Determining which one of the four is the underlying logical form lies beyond the aims of this discussion.

$$\text{ATOM}(B) \wedge \text{contain}(e, x, B)].$$

Despite the formal difference, both formulas express the weak reading, because the event summation is done across boxes in both cases, and the only difference is the atomicity of the theme (the books).

matrix-level	$\lambda e'$ -level	λe -level	corresponding reading
EXH	EXH	EXH	strong
		\emptyset	weak
	\emptyset	EXH	strong
		\emptyset	weak
\emptyset	EXH	EXH	strong
		\emptyset	weak
	\emptyset	EXH	strong
		\emptyset	literal

Table 1.1: Set of candidate meanings in Zweig’s approach

1.3 Presuppositional Exhaustification approach

Following Bassi, Del Pinal, and Sauerland 2021, the negation of non-weaker alternatives is part of the presupposition of a proposition, rather than part of its assertive content. For a proposition p , the exhaustification operator **PEX** has the following (provisional) output:

$$(33) \quad \mathbf{PEX}(p) = \begin{cases} \text{assertion: } p \\ \text{presupposition: } \bigwedge_{\substack{q \in \text{ALT}^*(p) \\ p \neq q}} \neg q \end{cases}$$

Thus, **PEX**(p) is true iff its assertive content and presuppositional content are true, false iff its assertive content is false, and undefined otherwise (i.e. in cases of presupposition failure):

$$(34) \quad \llbracket \mathbf{PEX}(p) \rrbracket = \begin{cases} 1 & \text{if } \llbracket p \rrbracket = 1 \wedge \left(\bigwedge_{\substack{q \in \text{ALT}^*(p) \\ p \neq q}} \neg q \right) = 1 \\ 0 & \text{if } \llbracket p \rrbracket = 0 \\ \# & \text{otherwise} \end{cases}$$

This definition can also be rewritten in terms of **EXH**:

$$(35) \quad \llbracket \mathbf{PEX}(p) \rrbracket = \begin{cases} 1 & \text{if } \llbracket \mathbf{EXH}(p) \rrbracket = 1 \\ 0 & \text{if } \llbracket p \rrbracket = 0 \\ \# & \text{otherwise, i.e. if } \llbracket p \rrbracket = 1 \wedge \llbracket \mathbf{EXH}(p) \rrbracket = 0 \end{cases}$$

The Strong Kleene semantics for the universal quantifier –with restrictor P and scope Q – is defined as follows:

$$(36) \quad \llbracket \text{Every } P \text{ is } Q \rrbracket = \begin{cases} 1 & \text{if } \forall x. \llbracket P \rrbracket(x) = 1 \rightarrow \llbracket Q \rrbracket(x) = 1 \\ 0 & \text{if } \exists x. \llbracket P \rrbracket(x) = 1 \wedge \llbracket Q \rrbracket(x) = 0 \\ \# & \text{otherwise} \end{cases}$$

We can now apply these definitions to “every box contains books” under its logical form (37)⁹:

$$(37) \quad \begin{aligned} & \llbracket \text{Every box } \lambda_x \mathbf{PEX}[x \text{ contains books}] \rrbracket \\ &= \begin{cases} 1 & \text{if } \forall x. \llbracket \mathbf{box} \rrbracket(x) = 1 \rightarrow \llbracket \mathbf{PEX}[\text{contains books}] \rrbracket(x) = 1 \\ 0 & \text{if } \exists x. \llbracket \mathbf{box} \rrbracket(x) = 1 \wedge \llbracket \mathbf{PEX}[\text{contains books}] \rrbracket(x) = 0 \\ \# & \text{otherwise} \end{cases} \\ &= \begin{cases} 1 & \text{if } \forall x. \llbracket \mathbf{box} \rrbracket(x) = 1 \rightarrow \llbracket \text{contains several books} \rrbracket(x) = 1 \\ 0 & \text{if } \exists x. \llbracket \mathbf{box} \rrbracket(x) = 1 \wedge \llbracket \text{contains books} \rrbracket(x) = 0 \\ \# & \text{otherwise} \end{cases} \\ &= \begin{cases} 1 & \text{if every box contains several books} \\ 0 & \text{if at least one box is empty} \\ \# & \text{otherwise} \end{cases} \end{aligned}$$

1.4 Homogeneity-based approach

Križ 2017 also relies logical trivalence, but for a purpose different from the **PEX** approach. The proposal in Križ 2017 is based on the phenomenon of homogeneity, whereby a sentence with a definite plural or a bare plural has its truth conditions in affirmative environments differ from the complementary of the truth conditions in negative environments. The “undefined” truth-value is introduced in order to account for the truth conditions in DE environments, while it was introduced to express presupposition failure in the **PEX** approach:

⁹The Presuppositional Exhaustification approach posits that **PEX** is applied locally. However, in an alternative approach that would apply **PEX** globally, it would in principle predict a weak reading.

$$\begin{aligned}
(38) \quad a. \quad \llbracket \text{Box } x \text{ contains books} \rrbracket &= \begin{cases} 1 & \text{if box } x \text{ contains several books.} \\ 0 & \text{if box } x \text{ contains no books.} \\ \# & \text{if box } x \text{ contains exactly one book.} \end{cases} \\
b. \quad \llbracket \text{Box } x \text{ doesn't contain books} \rrbracket &= \begin{cases} 1 & \text{if box } x \text{ contains no books.} \\ 0 & \text{if box } x \text{ contains several books.} \\ \# & \text{if box } x \text{ contains exactly one book.} \end{cases}
\end{aligned}$$

The application of a unary quantifier Q (such as the universal quantifier “every box”) to a trivalent predicate P (here, “contains books”) follows this rule, reproduced from Križ 2017:

- (39) a. Let P^1 be that predicate which is just like P except that it is true of all the individuals where P is undefined, and P^0 that predicate which is just like P except that it is false of all the individuals where P is undefined.
- b. If $Q(P^1)$ and $Q(P^0)$ have the same truth-value, then $Q(P)$ has that truth-value.
- c. Otherwise, $Q(P)$ is undefined.

P^1 and P^0 are two ways of resolving undefined statements, by conflating undefinedness into either truth or falsity conditions. Applying them to our predicate yields:

$$\begin{aligned}
(40) \quad a. \quad \llbracket x \text{ contains books} \rrbracket^1 &= \begin{cases} 1 & \text{if } x \text{ contains at least one book.} \\ 0 & \text{if } x \text{ contains no books.} \end{cases} \\
b. \quad \llbracket x \text{ contains books} \rrbracket^0 &= \begin{cases} 1 & \text{if } x \text{ contains several books.} \\ 0 & \text{if } x \text{ contains one or fewer books.} \end{cases}
\end{aligned}$$

We then examine the truth conditions of $Q(P^1)$ and $Q(P^0)$:

$$\begin{aligned}
(41) \quad a. \quad \llbracket \text{Every box } \lambda_x [x \text{ contains books}]^1 \rrbracket &= \begin{cases} 1 & \text{if each box contains at least one book.} \\ 0 & \text{else (i.e. if there is at least one empty box).} \end{cases} \\
b. \quad \llbracket \text{Every box } \lambda_x [x \text{ contains books}]^0 \rrbracket &= \begin{cases} 1 & \text{if each box contains several books.} \\ 0 & \text{else (i.e. if at least one box contains zero or one book).} \end{cases}
\end{aligned}$$

Cases in which $Q(P^1)$ and $Q(P^0)$ have the same truth-value are:

- if each box contains several books (truth conditions);

- if there is at least one empty box (falsity conditions).

Therefore:

$$(42) \quad \llbracket \text{Every box } \lambda_x [x \text{ contains books}] \rrbracket = \begin{cases} 1 & \text{if each box contains several books.} \\ 0 & \text{if there is at least one empty box.} \\ \# & \text{otherwise} \end{cases}$$

1.5 Interim summary: predictions of the different theories

Distinguishing between three readings is the minimal level of refinement needed to differentiate the theories discussed.

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.

The relative logical strengths of these readings are: strong > weak > literal. This has an important methodological consequence which influenced our design presented in Section 2.3, namely the fact that it is impossible to test a situation where *only* the strong reading or *only* the weak reading is true (the readings that they entail will also be true).

Let us now summarize, for each approach presented, the set of readings supported by a sentence with a bare plural in the scope of a universal quantifier.

Both bivalent approaches we presented posit an exhaustivity operator. Since there is a choice at each scope site to apply or not to apply the operator, the literal reading is always predicted in bivalent approaches, at least among candidate meanings, and corresponds to the case where the operator is never applied. After generating candidate meanings, the approach based on Zweig and Ivlieva posits an extra principle (29) to select the ultimate reading. The HOI approach does not posit a systematic selection principle, and thus predicts that three readings are generally possible. The availability of the weak reading is crucial, as it is only predicted in the HOI approach.

Under trivalent theories, the sentence is undefined in situations where the literal reading is true but the strong reading is false. However, it could be contextually considered “true enough” when not false: we indicate this with ‘(#)’.

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

Chapter 2

Experiments on plurals in the scope of universal quantifiers

2.1 Experimental studies in previous literature

2.1.1 Context-sensitivity of bare plurals

The experiments from Jiang and Sudo 2023 have aimed at comparing empirical support for the implicature-based approaches and the homogeneity-based approach, by examining context-sensitivity and using non-quantified as well as quantified sentences. The results show that truth-value judgments align more closely with implicature-based approaches. Our experiments include three key modifications compared to Jiang and Sudo 2023:

1. The stimuli in Jiang and Sudo 2023 include only one type of mixed scenario, where the proportion of single-object referents vs. multiple-object referents is 50/50. This is the case, for instance, of Figure 1, where 5 out of 10 boxes contain a single book and equally many boxes contain several books. One parameter absent from Jiang and Sudo 2023 that we will be using in my experiments is to introduce a more fine-grained distribution of multiple-object referents among the referents that are quantified over, making it possible to observe gradient effects.
2. Our experimental conditions do not manipulate context and we focus on explaining gradient effects in truth-value judgments. These effects were noticeable to a limited extent in the results of Jiang and Sudo 2023, but only across two conditions. The authors also do not discuss these effects, focusing rather on the symmetry of context-sensitivity with respect to polarity.
3. Jiang and Sudo 2023 introduce the implicature-based predictions as if they were all the same across different theories, but the previous chapter has shown that this is not the case.

2.1.2 Gradient effects in production

In Enguehard 2024, English speakers completed a production task regarding their choice between singular and plural indefinites under negation. Participants learned a rule that they had to formulate in English, after being shown a set of cards containing abstract symbols and sorted into categories of *valid* and *invalid* cards with immediate feedback (Figure 2.1).

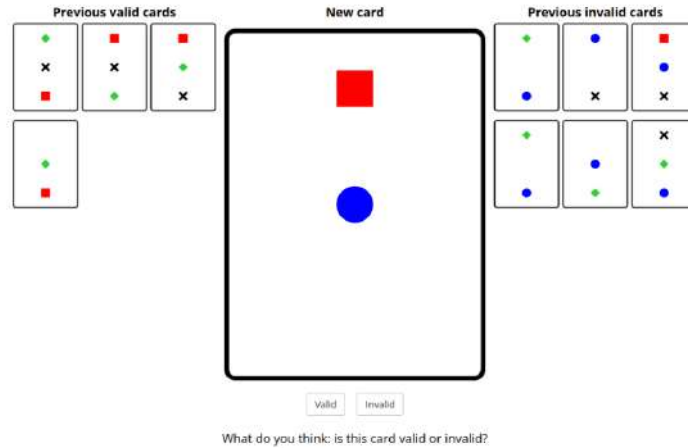


Figure 2.1: Example of a trial, reproduced from Enguehard 2024.

Participants then had to complete the sentence “The card is valid when...”. The setup prompted them to use negated indefinites, e.g. “The card is valid when it has no blue circles”. The actual rule was: a card is valid when it does not have any blue circles. There were different probability distributions of the number of referents (here, blue circles) across invalid cards: only singular or only plural referents (i.e. cards having a singular blue circles or multiple blue circles), a large majority of singular or plural, or a balanced 50/50 situation. The choice of number marking was observed to be influenced in a gradient manner by the probability distribution. In singular-only or plural-only conditions, almost only singular or plural indefinites are used. In intermediate conditions, the proportion of plural productions increases with the proportion of plural referents.

We argue that these observations could, to some extent, be predicted by combining presuppositional exhaustification (Ahn, Saha, and Sauerland 2020) and a constraint on informativity adapted from Doron and Wehbe 2022. Following the definition (35) from the **PEX** approach, the bare plural “circles” presupposes “zero or at least two circles”. This is because undefinedness corresponds, on the one hand, to presupposition failure, and on the other hand, to situations where the literal meaning is true but not the meaning strengthened with **EXH** (for bare plurals, this is when *at least one* is true but not *several*, i.e. situations of *exactly one*). By identifying these two undefinedness conditions, we know that presupposition failure occurs when *exactly one* is false. Hence, the presupposition of a

bare plural is “zero or several”.

We aim at explaining how the use of (43) varies with the probability distribution. The sentence in its original form does not have a very transparent meaning, as it contains a generic “the card” and a “when”-clause.

(43) The card is valid when it has no blue circles.

We will reformulate (43) as:

(44) If the card has no blue circles, then it is valid.

Following the principle of presupposition projection in a conditional sentence, (44) presupposes that the card has zero or several blue circles. The generic expression “the card” poses some issues, because if it means *any card*, then we might expect to encounter presupposition failure as soon as there is one card (valid or invalid) that has been shown and has exactly one blue circle. For simplicity, we will take “the card” to be a referential expression meaning *the next card that will be shown*.

To this, we add a constraint on informativity adapted from (45), which is repeated from Doron and Wehbe 2022. We will also need a probabilistic definition of informativity given in (46):

(45) **Post-Accommodation Informativity (PAI):** A sentence S_p (presupposing p) can be uttered felicitously only if S_p is informative w.r.t. the QUD and common ground after presupposition accommodation.

(46) A proposition p is informative with respect to a common ground C if $\mathbb{P}(p|C) < 1$.

To illustrate the reasoning using (45) and (46), imagine that all valid cards that have been shown have no blue circles and all invalid cards have exactly one blue circle. This context creates a certain prior, namely the prior that a card with several blue circles is not likely to appear. If C denotes the common ground before presupposition accommodation, then the common ground C' after accommodation is the intersection of C and the set of worlds where the next card has zero or several blue circles. Given the priors in C , it seems much more probable for the next card to have zero blue circles than several blue circles, because no card among those already shown has several blue circles. Therefore, C' is the set of worlds where the next card has no blue circles. This is exactly the assertive content of (43), which becomes uninformative according to the PAI constraint. (45) and (46) together seem to predict the infelicity of (43) in the situation described.

Now consider a situation where all valid cards have no blue circles and where among invalid cards, 90% have only one blue circle and 10% have several blue

circles. Intuitively, (44) feels not outright uninformative but only a little informative. To account for this, we can incorporate into (46) a gradient notion of informativity:

- (47) Given a proposition p and a common ground C , the closer $\mathbb{P}(p|C)$ is to 0, the more informative p is. Conversely, the closer $\mathbb{P}(p|C)$ is to 1, the less informative p is.

In this scenario, the next card is presupposed to have zero or several blue circles, but has an expected probability of $\frac{10}{11}$ to have zero blue circles, against an expected probability of $\frac{1}{11}$ to have several blue circles. The common ground after accommodation C' is such that $\mathbb{P}(p|C) = \frac{10}{11}$ given the priors. Considering (47), this explains why (44) is not informative.

However, there are several limitations to our tentative reasoning. Firstly, we made the simplifying assumption that “the card” was referencing the next card, while the expression is supposed to be generic. Secondly, projection of the presupposition of a generic expression may prove less straightforward than in the case of referential expression. Thirdly, our arguments do not extend to contexts where gradient effects are hard to explain in terms of priors and triviality after presupposition accommodation. Thus, we conducted a production study presented thereafter, to observe if English production choices display gradient effects in mixed scenarios where they cannot be explained in the same way as above.

2.2 Production study

We tested the hypothesis that the higher the proportion of unique-object boxes in the picture, the greater the proportion of participants who will use a singular indefinite to complete the universally quantified sentence “Every box contains...”, and conversely for bare plurals. We expected gradient effects, such that the proportion of unique-object boxes will positively correlate with the proportion of singular indefinites used and negatively correlate with the proportion of bare plurals used.

2.2.1 Methods

Participants

We tested 250 adult participants recruited through Prolific (mean age 39.1; age range 19-77; 157 females). They were paid £0.3 for their participation. The sample size was chosen to ensure that we would collect approximately 15 to 20 valid responses per condition across the 11 experimental conditions.

Participants had to fulfill the following eligibility criteria:

1. report English to be their ‘first language’, ‘primary language’, and ‘earliest language in life’;
2. be located in the US or in the UK (to maximize the likelihood of being active English speakers).

Procedure and stimuli

A trial consists of a picture and an unfinished sentence to be completed. The picture consisted of 10 boxes, each box containing between 1 and 4 objects, with the number of objects possibly differing between boxes. The experimental condition is defined as the number of boxes containing a unique object. Each condition is instantiated 11 times, with 11 different objects all familiar to participants (apples, bikes, birds, books, chairs, flowers, houses, pencils, rabbits, stars, trees). Across all possible stimuli, the number of boxes containing a unique object will vary from 0 to 10.

Each participant completed one trial. They were asked to complete the sentence “Every box contains...”, based on the picture shown. Here is an example of a trial:

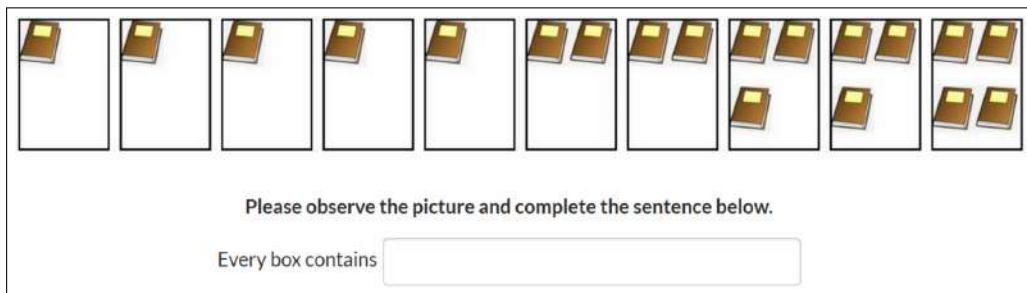


Figure 2.2: Example stimulus of the production study

On the second page (from which participants cannot navigate back), they answer a follow-up question: “How many Xs were there in the box(es) with the fewest Xs?”, where X is replaced by the object name.

The final page included an attention check.

2.2.2 Analyses

Participants who completed the full experiment were excluded if:

1. they failed the attention check (6 participants);
2. they answer the follow-up question incorrectly (19 participants).

For the remaining participants, we manually classified the answers into these categories:

- simple singular noun phrases (*a NP*);
- bare plural noun phrases (*NPs*);
- expressions equivalent to *one or more*, that could be considered as expressing *general number* (those included “one or more Xs”, “one, two, or three Xs”, “at least one X”...).

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. This model was compared to a null model without the predictor, using a likelihood ratio test (LRT). We expected a positive slope in the logistic model and a significant LRT p -value. We corrected for multiple comparisons using the Holm-Bonferroni method.

We also performed the LRT excluding the extreme conditions ($n = 0$ and $n = 10$, where n is the number of boxes with a single object). This is because we aim at observing gradience across mixed conditions, and the extreme conditions are not mixed.

2.2.3 Results

As shown in Figure 2.3, higher proportions of unique-object boxes increased the use of *a NP*, while bare plural use declined. A logistic regression predicting singular use significantly outperformed a null model:

- all conditions: $\chi^2(1) = 29.92, p < 0.001$;
- excluding extreme conditions: $\chi^2(1) = 9.20, p = 0.002$.

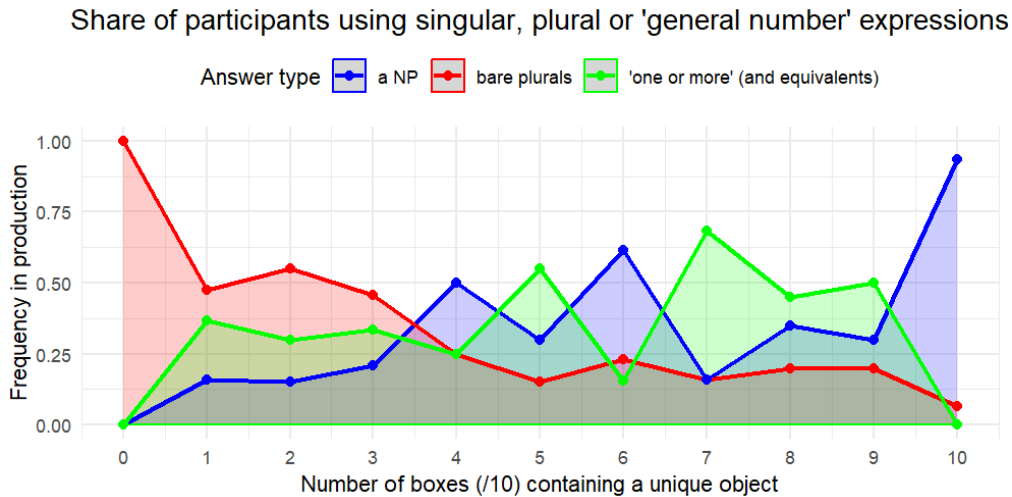


Figure 2.3: Production results in English

2.2.4 Discussion

In response to our speculation in Paragraph 2.1.2, gradient effects in this production study cannot be explained in terms of priors¹. In line with our hypothesis, we observed gradient effects of *a NP* used in production, positively correlating with the proportion of unique-object boxes. Likewise, the proportion of bare plurals positively correlated with the proportion of boxes containing several objects. However, the increase of *a NP* proportion was not perfectly regular. To a lesser extent, the decrease of bare plurals proportion was not perfectly regular either. This could be due to accidental variations, linked to the small number of participants (10 to 15) in each condition.

2.3 Comprehension study on bare plurals

Following the English production study, we conducted a pilot comprehension study using the same natural objects stimuli. Each participant completed a single trial and provided a continuous judgment of the sentence “Every box contains [bare plural]” using a cursor. The results showed little evidence of gradient judgments, with mean scores consistently high across all conditions (above 50 on a 1-100 scale). These findings made us introduce several modifications to the design of comprehension experiments:

1. We switched from a between-subjects to a within-subjects design, so that each participant was exposed to all conditions in randomized order.
2. We added literally false conditions, in which at least one box was empty. This is to increase contrast between conditions and to make it possible to observe gradience even among false cases.
3. We reduced the number of boxes from 10 to 4 in order to have enough mixed distributions (needed to observe gradience) while keeping the total number of stimuli reasonable.
4. In the pilot study, the uniformly-singular condition yielded a mean rating above 60 out of 100. This unexpectedly high rating may be due to a cumulative reading, whereby participants judged “Every box contains [bare plural]” as true even when each box contained only one object, as long as there were several objects in total. This interpretation is a confound, as our target interpretation depends on multiplicity relative to each entity that is quantified over entity (here, boxes). Because cumulative readings

¹Émile Enguéhard (p.c.) pointed out to us that for non-contextualized quantification over plural indefinites, gradience in production could be explained by increased anaphoric potential: the more boxes contain several objects, the more speakers are inclined to use a bare plural, as it enables anaphoric reference to a greater portion of those objects than a singular indefinite would.

are less accessible with “each” than with “every,” we replaced “every” with “each” in the stimuli sentences.

5. We compared mean judgments across different natural objects used in the stimuli and found that some items, like pencils and trees, consistently received lower ratings than other items, like flowers. This is possibly due to differences in perceived countability. To eliminate this confound, we replaced natural objects with geometric shapes in subsequent experiments.

The main novelty of our experimental design lies in its ability to capture gradience within a single level of reading, something not achieved in previous literature. This was made possible by including more than two literally false conditions, and more than two ‘truly mixed’ conditions, i.e. conditions that satisfy the weak reading but not the strong reading. This was not the case in previous work (Chemla and Spector 2011; Stateva, Andreetta, and Stepanov 2016; Jiang and Sudo 2023), where experiments did have ‘truly mixed’ conditions but had only two. Although all three studies reported differences in judgments between these conditions, neither accounted for gradient effects.

2.3.1 Methods

Participants

We tested 200 adult participants recruited through Prolific (mean age 41.7; age range 19-77; 119 females). They were paid £0.75 for their participation. The sample size was based on power analyses from the pilot study mentioned above, yielding a statistical power over 80%.

Participants had to meet the same eligibility criteria as in the previous experiment and must not have participated in the previous experiment.

Procedure and stimuli

A trial consisted of a picture and a sentence. The sentence followed a fixed structure: “Each box contains [bare plural]”, where the [bare plural] referred to the geometric shape present in the image (circles, triangles, or squares). The picture consisted of four boxes containing between 0 and 4 geometric shapes.

There were two types of conditions in the experiment:

- **Literal truth conditions:** a sentence was considered literally false if at least one box was empty (4 conditions), and literally true otherwise (5 conditions).
- **Experimental conditions:** a box containing multiple shapes was called a *strong verifier*. Conditions were labeled using the number of strong verifiers in the picture and the strongest reading they satisfied, namely:

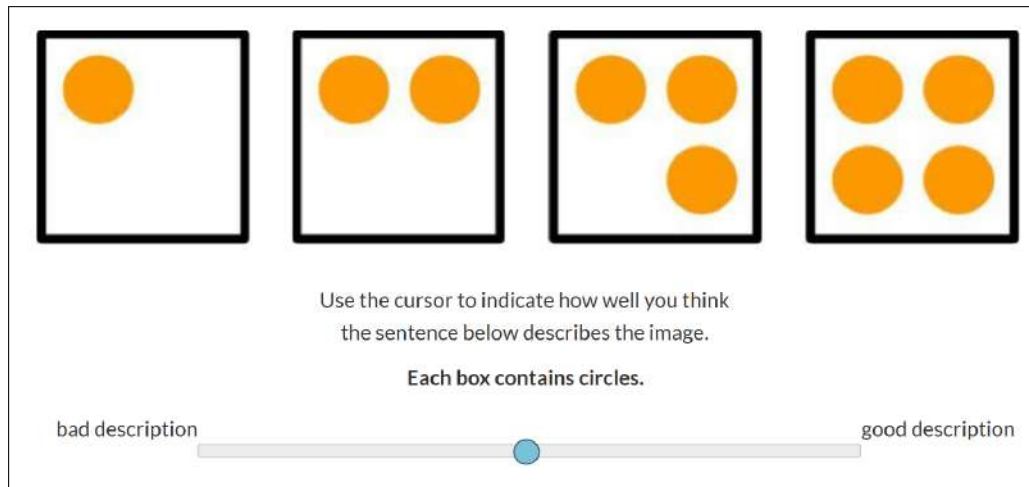


Figure 2.4: Example stimulus of the comprehension study with bare plurals

- **FALSE:** no reading is true (at least one box is empty);
- **LITERAL:** only the literal reading is true (every box contains at least one shape);
- **WEAK:** the literal reading and the weak reading are true, but not the strong reading (every box contains at least one shape and it is not the case that every box contains at least two shapes);
- **STRONG:** all readings are true (every box contains at least two shapes).

Each trial was accompanied by the same instruction: “Use the cursor to indicate how well you think the sentence below describes the image.” The cursor moved on a continuous scale ranging from “bad description” (left extremity) to “good description” (right extremity). Participants’ responses were recorded as integers between 1 and 100, but the numeric rating was not visible to participants.

Each participant completed 27 randomized trials, with each condition appearing three times, using three different geometric shapes in three different colors. The exact pairing of shapes and colors is detailed in the preregistration document (Appendix B).

The final page included an attention check.

2.3.2 Analyses

Participants who completed the full experiment were excluded if:

1. they failed the attention check (2 participants);
2. they did not rate every trial of the condition **STRONG-4** higher than every trial of the condition **FALSE-0** (10 participants).










stimuli pictures (examples with circles)	label for condition	readings
	STRONG-4	true for STRONG+WEAK+LITERAL readings
	WEAK-3	true for WEAK+LITERAL readings
	WEAK-2	true for WEAK+LITERAL readings
	WEAK-1	true for WEAK+LITERAL readings
	LITERAL-0	true for LITERAL reading
	FALSE-3	false for LITERAL reading
	FALSE-2	false for LITERAL reading
	FALSE-1	false for LITERAL reading
	FALSE-0	false for LITERAL reading

Figure 2.5: Table of stimuli for bare plurals, with corresponding readings

We defined 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)

We subset the literally true conditions (**LITERAL**, **WEAK**, **STRONG**) and fit a linear mixed-effects model predicting responses as a function of c_{vrf} , with random intercepts and slopes by participant.

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

In case of convergence issues or singular fit, the random slope would be removed. We compared this model to a null model containing only the intercept using a

likelihood ratio test (LRT). We expected a positive slope in the linear model and a significant LRT p -value.

We also performed the LRT on the **WEAK** conditions, because the set of **WEAK** conditions represent mixed scenarios, contrary the conditions **LITERAL-0** and **STRONG-4**. Crucial to our analysis is whether gradient effects are found across **WEAK** conditions alone.

As an exploratory analysis (not preregistered), we conducted model comparisons across all 9 conditions using the Bayesian information criterion (BIC) and the Akaike information criterion (AIC) to identify the best-fitting combination of predictors, among the $2^4 = 16$ possible combinations. Our primary claims will be based on BIC results rather than AIC results, because BIC is more appropriate when testing between theoretically motivated predictors, as is the case here. BIC gives more penalty for model complexity, compared to AIC. We will still generate AIC-based rankings as an exploratory follow-up.

2.3.3 Results

Results are shown in Figure 2.6. The top graphs are histograms of the scores in each condition.

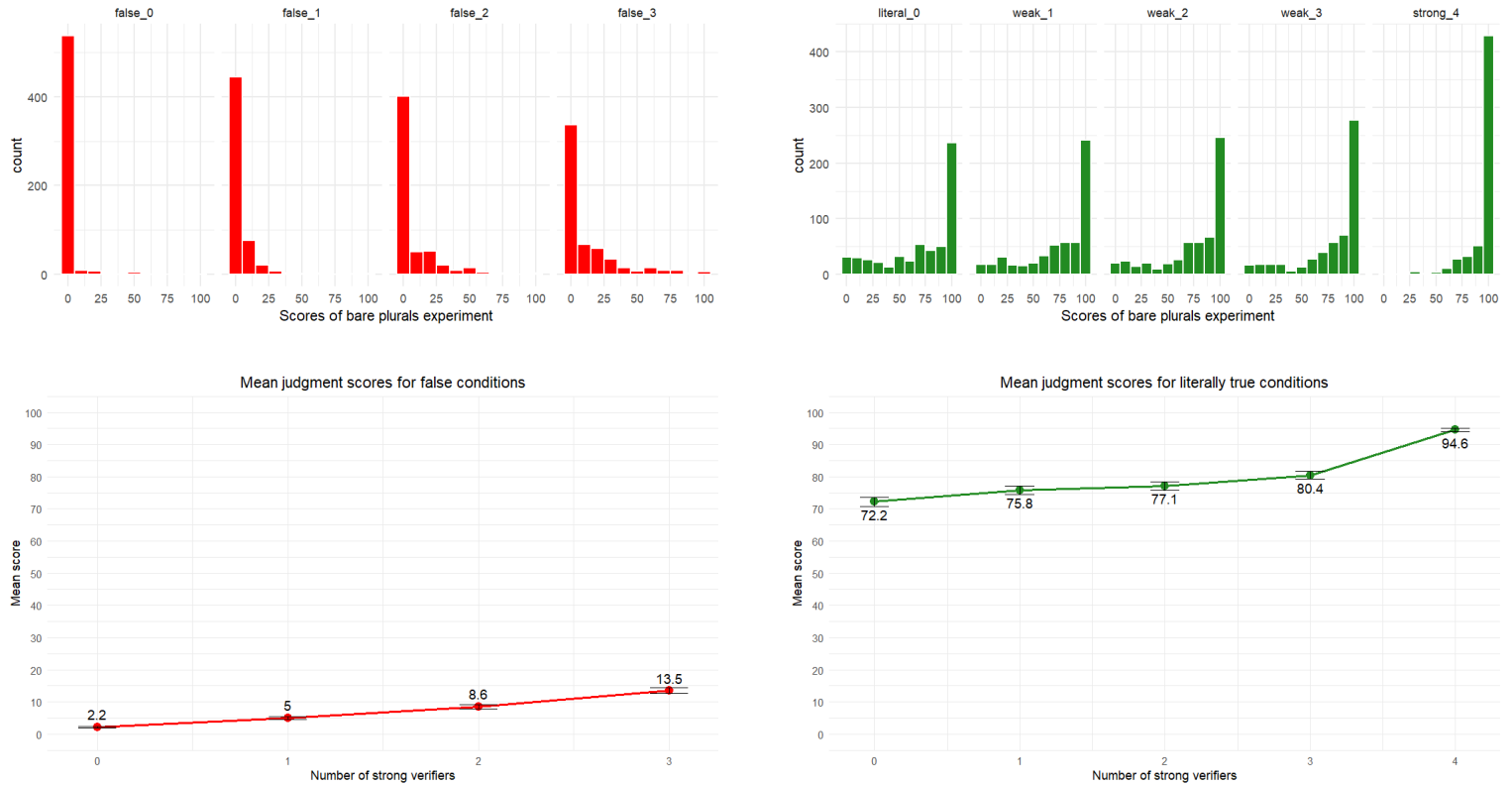


Figure 2.6: Results of the comprehension task with bare plurals

While fitting a mixed-effects model to the subset of literally true conditions, the random slope was removed due to failure of convergence. The model fit on literally true conditions (c_{vrf} as predictor) significantly outperformed the null model ($\chi^2(1) = 395.01, p < 10^{-15}$)². As expected, the effect of c_{vrf} was positive ($b = 4.93, SE = 0.24, t = 20.64$).

A gradient effect was also found within **WEAK** conditions alone ($\chi^2(1) = 18.75, p < 10^{-4}$), where the random slope allowed model convergence and was kept.

The best-fitting model across all 9 conditions was:

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

Table 2.1 reports statistical details of the best model.

Predictor	Estimate	Standard error	<i>t</i> -value
(Intercept)	2.561	1.049	2.441
c_{vrf}	3.180	0.256	12.413
c_{lit}	69.076	0.573	120.579
c_{str}	10.197	1.109	9.192

Table 2.1: Fixed effects of the best-fitting linear mixed-effects model (bare plurals experiment)

The second best-fitting model was:

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

with $\Delta BIC = 8$ between the best two models. This difference is considered strong evidence in favor of the better-ranked model, following standard guidelines (e.g. Raftery 1995).

The full model rankings using BIC and AIC are presented in Table 2.2. Both criteria yielded the same order of ranking.

2.3.4 Discussion

Our results showed gradient effects driven by c_{vrf} in both true and false conditions and, crucially, also within **FALSE** and within **WEAK** conditions. Gradiance across conditions of the same reading is evidence for the status of gradiance as a factor of its own (confirmed by the model comparisons), and not as a by-product of increased judgment scores when one more reading is made true.

²The p -values reported have been corrected for multiple comparisons using the Holm-Bonferroni method.

Rank	Model formula	BIC	AIC
1	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{str}} + (1 \mid \text{participant})$	44940	44901
2	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	44948	44902
3	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + (1 \mid \text{participant})$	45015	44980
4	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{weak}} + (1 \mid \text{participant})$	45019	44983
5	$\text{response} \sim c_{\text{lit}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	45058	45018
6	$\text{response} \sim c_{\text{lit}} + c_{\text{str}} + (1 \mid \text{participant})$	45083	45051
7	$\text{response} \sim c_{\text{lit}} + c_{\text{weak}} + (1 \mid \text{participant})$	45353	45320
8	$\text{response} \sim c_{\text{lit}} + (1 \mid \text{participant})$	45449	45423
9	$\text{response} \sim c_{\text{vrf}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	48255	48216
10	$\text{response} \sim c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	48500	48467
11	$\text{response} \sim c_{\text{vrf}} + c_{\text{weak}} + (1 \mid \text{participant})$	48600	48568
12	$\text{response} \sim c_{\text{weak}} + (1 \mid \text{participant})$	48644	48617
13	$\text{response} \sim c_{\text{vrf}} + c_{\text{str}} + (1 \mid \text{participant})$	51677	51644
14	$\text{response} \sim c_{\text{str}} + (1 \mid \text{participant})$	51707	51680
15	$\text{response} \sim c_{\text{vrf}} + (1 \mid \text{participant})$	52088	52062
16	$\text{response} \sim 1 + (1 \mid \text{participant})$	52556	52537

Table 2.2: Model comparison by BIC and AIC (bare plurals experiment)

We observed a qualitative shift in judgments between the **WEAK** conditions and the **STRONG-4** condition, but only a quantitative increase (driven only by c_{vrf}) from **LITERAL-0** to **WEAK-1**. This asymmetry suggests that the weak reading may not be accessed in comprehension, at least in the absence of a context that could make the weak reading relevant. This seems to favor approaches that do not predict a weak reading.

We also noted that in a uniformly-singular situation, the mean judgment was over 70% in favour of “Each box contains [bare plural]” being a “good description” of the situation. We were surprised by such a high rating and proceeded to verify whether this judgment could have resulted from a marginal cumulative interpretation (supposedly less accessible with “each” than with “every”, but maybe still present). This motivated the follow-up experiment presented thereafter.

2.4 Comprehension study: cumulativity of different plural expressions

2.4.1 Methods

Participants

In a pilot study, we tested between 20 and 25 participants per condition and examined three interaction contrasts comparing the effect of box number (1 vs. 4) across sentence types. Based on the pilot data, a power analysis indicated that 80 participants per condition would yield approximately 80% power to detect all three interaction effects. To allow for exclusions and ensure conservative coverage, we recruited 100 participants per condition. Thus, we tested at total of 600 adult participants recruited through Prolific (mean age 42.6; age range 18-80; 328 females). They were paid £0.15 for their participation.

Participants had to meet the same eligibility criteria as in the previous experiments and must not have participated in the previous experiments.

Procedure and stimuli

A trial consisted of a picture paired with a sentence. The sentence followed a fixed structure: “[Each/The] box contains [plural expression]”. The image consisted of either one box or four boxes, each containing exactly one circle.

The experiment crossed two factors, yielding six conditions:

- **F1. Plural expression:** *several NPs*, *some NPs*, or bare plural;
- **F2. Picture type:** either a single box (sentence began with “The box”), or four boxes (sentence began with “Each box”).

Each participant completed exactly one trial. Assignment to condition was randomized, resulting in approximately 100 participants per condition. The rest of the design was identical to the previous experiment.

2.4.2 Analyses

Participants who did not complete the full experiment were excluded from the analyses.

To assess the impact of sentence type and picture type on acceptability ratings, we fitted a linear model predicting scores from sentence type (bare plural, *some NPs*, or *several NPs*), picture type (1-box or 4-boxes), and their interaction. This allowed us to evaluate both the main effects of each factor and whether the influence of visual context differed by sentence type.

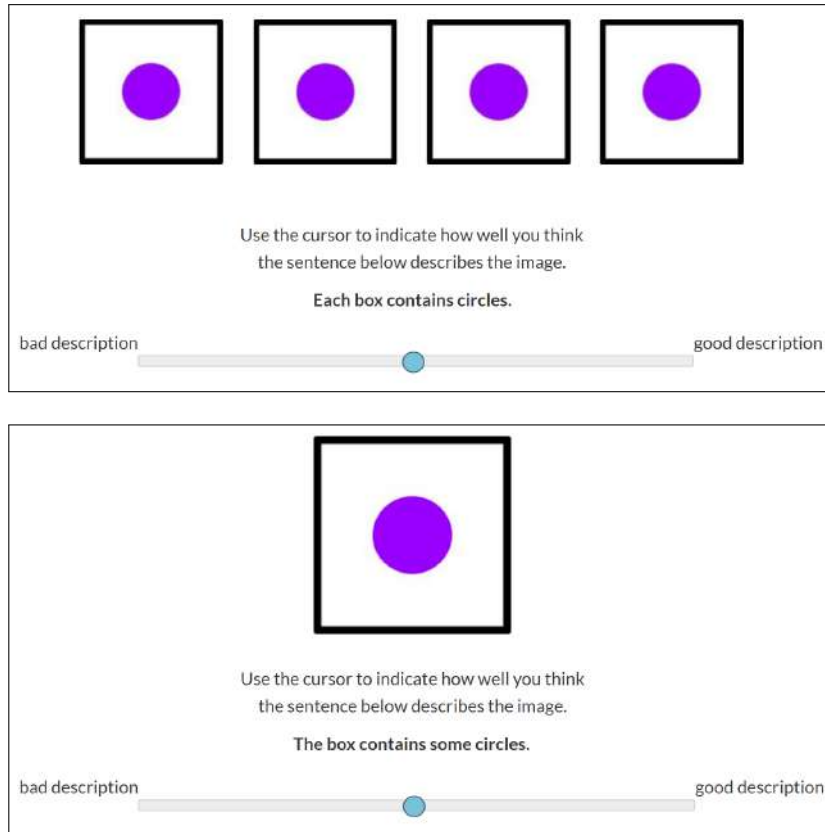


Figure 2.7: Example stimuli of the cumulativity experiment

To explore any interaction between sentence type and image type, we conducted three pairwise interaction contrasts. These contrasts compared the magnitude of the box effect (i.e. the difference between 1-box and 4-box ratings) across sentence types. Each contrast was evaluated using a Wald t -test based on estimated marginal means obtained via the `emmeans` package in R.

2.4.3 Results

The mean ratings for each condition are displayed in Table 2.3. Across all sentence types, participants gave higher ratings to sentences paired with 4-box images than to those paired with 1-box images.

	Bare plural	<i>Some NPs</i>	<i>Several NPs</i>
1-box	32.8	18.9	3.4
4-boxes	75.8	46.8	14.5

Table 2.3: Mean ratings (scale 1-100) by sentence type and image type

The linear model revealed a significant main effect of picture type: across all sentence types, sentences paired with 4-boxes images received higher acceptability ratings than those paired with 1-box images.

Importantly, the interaction between sentence type and picture type was also significant. Results for interaction contrasts are presented in Table 2.4.³

Sentence contrast	Estimate	SE	df	t-value	p-value
BP vs. <i>several</i>	−31.997	6.095	594	−5.250	< 10 ^{−8}
BP vs. <i>some</i>	−15.203	5.933	594	−2.563	0.001
<i>several</i> vs. <i>some</i>	16.795	6.091	594	2.757	0.001

Table 2.4: Pairwise interaction contrasts of picture type between sentence types

2.4.4 Discussion

Our results support the hypothesis that participants are sensitive to a possible cumulative interpretation when multiple boxes are shown in total.

The effect of *some NPs* across image types was smaller in magnitude than the effect of bare plurals. Furthermore, the mean rating of the 1-box scenario was significantly lower with *some NPs* than with bare plurals. This is unaccounted for, as existing theories do not predict different availability levels of cumulative readings between bare plurals and *some NPs*. The results also showed that *several NPs* resists cumulative interpretations. More theoretical investigations are necessary to explain why the ability to license cumulative interpretations may partially depend on the type of indefinite plural expression.

Following this experiment, we conducted two more comprehension studies with the same design and analyses as the bare plurals experiment, replacing bare plurals with *several NPs* and *some NPs*.

2.5 Comprehension study on *several NPs*

With *several NPs*, there are no levels of readings involved, as multiplicity is encoded in the quantifier’s denotation. As only one reading can be true, this is helpful for testing a hypothesis related to the source of gradience: it could be that gradient effects reflect proximity to the closest situation that makes a certain reading true (in the case of this experiment, it can only be the strong reading). Thus, we focused on observing whether gradience could still be found across conditions in which “Every box contains [several NPs]” is false, that is, 8 conditions out of 9.

³All *p*-values reported have been corrected for multiple comparisons using the Holm-Bonferroni method.

Note that, under this hypothesis, we did expect gradient effects to arise with *several NPs*, because of what we observed within **FALSE** cases in the bare plurals experiment. Gradience across the four literally false conditions could only be due to two factors: more plausibly, the distance to the closest situation making the literal reading true; less plausibly, the total number of geometric shapes in the picture.

2.5.1 Methods

Methods were mostly identical to those of the bare plurals experiment (Section 2.3), with the following differences:

1. **Participants.** We tested 70 adult participants recruited through Prolific (mean age 37.1; age range 19-72; 39 females). Participants were paid £0.60 for their participation. They had to meet the same eligibility criteria as in the previous experiments and must not have participated in the previous experiments.
The sample size was based on power estimates using the data collected through the bare plurals experiment. A sample size of 70 yielded a statistical power over 90%.
2. **Stimuli.** Sentences in the trials followed the fixed structure: “Each box contains [several NPs]”. The rest of the design remained the same as before.
3. **Conditions.** There were, again, two types of conditions, slightly different from those of the previous experiment:
 - **Literal truth conditions:** a sentence is literally true (in what follows, we will also call it *strong*) if all boxes contain several shapes (1 condition), and literally false otherwise (8 conditions).
 - **Experimental conditions:** as before, a box containing multiple shapes was called a *strong verifier*. Conditions were labeled based on the reading they satisfy (**FALSE**, **STRONG**), followed by the number of empty boxes, then by the number of strong verifiers in the image. The two numbers reflect the two ways of falsifying *several NPs*, i.e. either an empty box or a box with a unique shape.

2.5.2 Analyses

As in the bare plurals experiment, participants who completed the full experiment were excluded if:

1. they failed the attention check (0 participants);

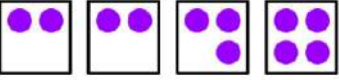
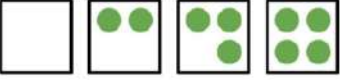
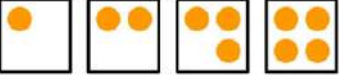
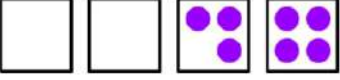


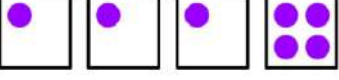


stimuli pictures (examples with circles)	label for condition	stimuli pictures (examples with circles)	label for condition
	STRONG-0-4		FALSE-1-3
	FALSE-0-3		FALSE-2-2
	FALSE-0-2		FALSE-3-1
	FALSE-0-1		FALSE-4-0
	FALSE-0-0		

Figure 2.8: Table of stimuli for *several NPs*

- they did not rate every trial of the condition **STRONG-0-4** higher than every trial of the condition **FALSE-4-0** (0 participants).

In this version, we only have one predictor, that is c_{vrf} , the number of strong verifiers. We subset the no-empty-box conditions (**FALSE-0-0**, **FALSE-0-1**, **FALSE-0-2**, **FALSE-0-3**, **STRONG-0-4**) and fit a linear mixed-effects model predicting responses as a function of c_{vrf} , with random intercepts and slopes by participant.

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

In case of convergence issues or singular fit, the random slope would be removed. We compared this model to a null model containing only the intercept using a likelihood ratio test. We did the same for the subset of at-least-one-empty-box conditions (**FALSE-4-0**, **FALSE-3-1**, **FALSE-2-2**, **FALSE-1-3**).

2.5.3 Results

Results are shown in Figure 2.9.

There were no convergence issues or singular fit for either model. The model fit significantly outperformed the null model for the subset of no-empty-box conditions ($\chi^2(1) = 232.31, p < 10^{-49}$) as well as for the subset of at-least-one-empty-box conditions ($\chi^2(1) = 380.01, p < 10^{-81}$)⁴. Tables 2.5 and 2.6 report

⁴The p -values reported have been corrected for multiple comparisons using the Holm-Bonferroni method.

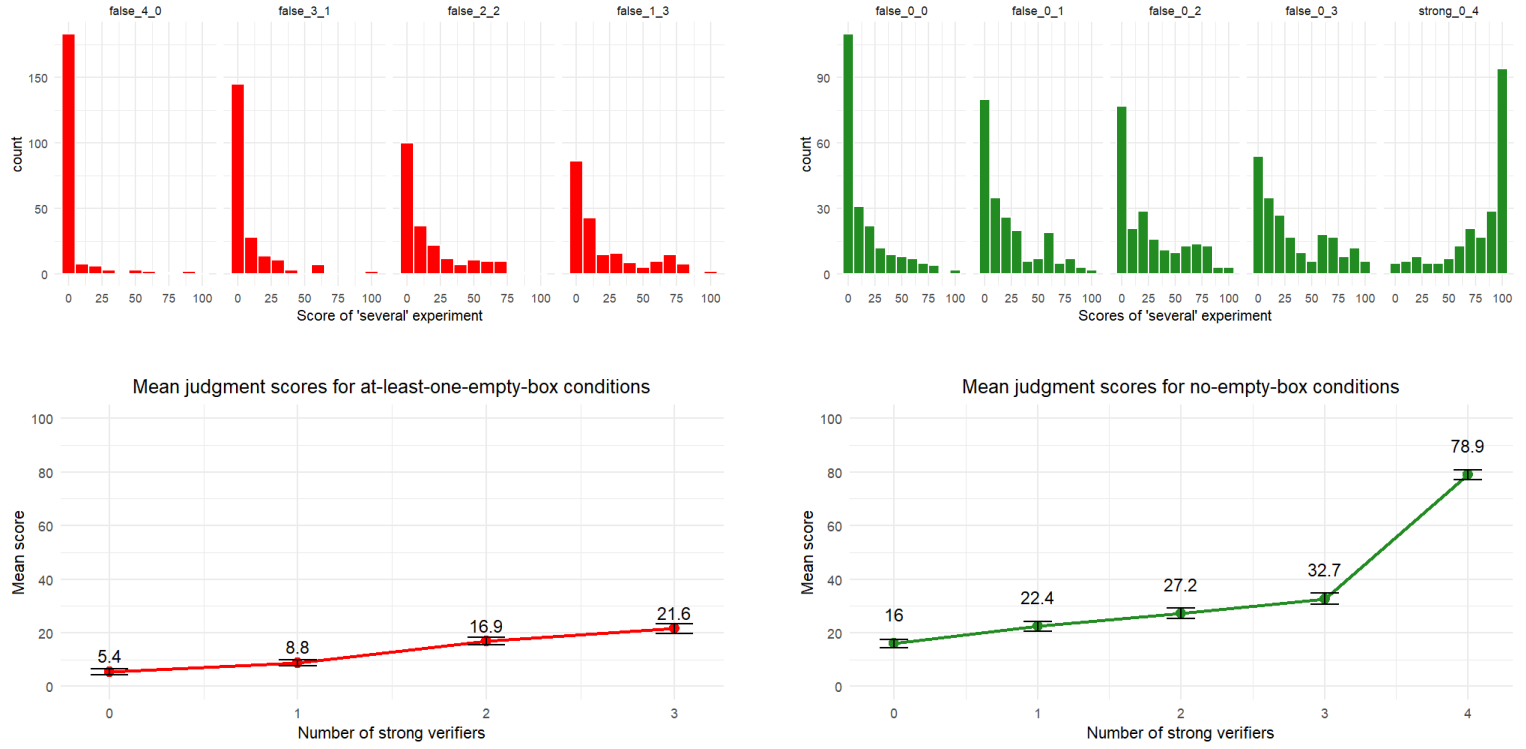


Figure 2.9: Results of the comprehension task with *several* NPs

statistical details of the two fitted models.

Predictor	Estimate	Standard error	<i>t</i> -value
(Intercept)	13.971	2.934	4.762
c_{vrf}	5.971	0.942	6.340

Table 2.5: Fixed effects for the no-empty-box model

2.5.4 Discussion

Again, we observed gradient effects driven by c_{vrf} within both subsets of conditions. This is in favor of our hypothesis that gradience reflects proximity to the closest situation making a certain reading true, at least in a case where there is only one possible true reading.

One important difference from the bare plurals experiment (as well as the *some* NPs experiment, cf. *infra*) was that mean judgments do not consistently increase with the order in which conditions are presented in the graphs. This is because the distinction between *at-least-one-empty-box* and *no-empty-box* conditions is somewhat arbitrary: the sentence with *several* NPs is literally false in

Predictor	Estimate	Standard error	t-value
(Intercept)	1.197	0.658	1.818
c_{vrf}	6.313	1.189	5.308

Table 2.6: Fixed effects for the at-least-one-empty-box model
(only false conditions)

all of them. This distinction between conditions was only introduced to reflect the separation between literally true and literally false conditions for bare plurals and *some NPs*.

Notably, the condition with the lowest mean rating among the *no-empty-box* conditions, **FALSE-0-0**, does not receive a higher rating than the highest-rated condition among the *at-least-one-empty-box* subset, **FALSE-1-3**. In the former, the total number of geometric objects is 4; in the latter, it is 9. There are at least two plausible reasons why **FALSE-1-3** is judged more acceptable despite the presence of an empty box. First, it contains a larger number of objects overall. Second, it is intuitively closer to the scenario that makes the sentence true, namely **STRONG-0-4**: the minimal change required to make **FALSE-1-3** true is smaller than the change required for **FALSE-0-0**.

A somewhat surprising observation is that the mean rating for **STRONG-0-4** is not extremely high, at 78.9. In comparison, the same condition received an mean rating above 90 in the bare plurals experiment⁵, possibly reflecting the common interpretation of “several” as denoting a number greater than or equal to 3, rather than 2. Another unexpected observation, based on metadata from Prolific, is that the median completion time for the *several NPs* experiment was nearly 50% longer than that of the bare plurals experiment (5’22” vs. 3’38”). Although the sample sizes differed substantially between the two experiments (200 vs. 70), this discrepancy in completion times may suggest greater hesitation or uncertainty among participants when evaluating *several NPs* statements.

2.6 Comprehension study on *some NPs*: continuous judgments

With *some NPs*, the range of available readings is the same as for bare plurals. As far as the theories presented in Chapter 1 are concerned, the predictions for bare plurals also apply to *some NPs*. However, as the cumulativity experiment showed, the mean rating for the uniformly-singular condition is 20 points lower with *some NPs* than with bare plurals. This suggests that there is more room for *some NPs* ratings to increase within literally true conditions. It is possible

⁵And also above 90 in the *some NPs* experiment, although it was conducted after the *several NPs* experiment.

that, in the bare plurals experiment, the combination of high baseline ratings and gradience within literally true conditions masked the qualitative shift from the weak to the literal reading. If so, *some NPs* should provide better conditions for observing this shift.

2.6.1 Methods

Methods were mostly identical to those of the comprehension study on bare plurals (Section 2.3), with the following differences:

1. **Participants.** After exclusions, we tested 200 adult participants recruited through Prolific (mean age 41.2; age range 18-73; 88 females). Participants were paid £0.60 for their participation. They had to meet the same eligibility criteria as in the previous experiments and must not have participated in the previous experiments.

The sample size was based on power estimates using the data collected through the bare plurals experiment. We conducted a bootstrap analysis where we required the difference in BIC between the best two models to be at least 4, to ensure stronger evidence in favor of the model with lower BIC. A sample size of 200 yielded a statistical power over 80%.

2. **Stimuli.** Sentences in the trials followed the fixed structure: “Each box contains [some NPs]”. The rest of the design remained the same as before.

2.6.2 Analyses

We conduct the same analyses as for the bare plurals experiment, including the model comparison using BIC that has been preregistered this time.

2.6.3 Results

Results are shown in Figure 2.10.

While fitting a mixed-effects model to the subset of literally true conditions, the random slope was removed due to failure of convergence. The model fit on literally true conditions (c_{vrf} as predictor) significantly outperformed the null model ($\chi^2(1) = 1052.9, p < 10^{-15}$)⁶. As expected, the effect of c_{vrf} was positive ($b = 8.98, SE = 0.25, t = 35.75$).

A gradient effect was also found within **WEAK** conditions alone ($\chi^2(1) = 65.19, p < 10^{-15}$), where the random slope allowed model convergence and was kept.

⁶The p -values reported have been corrected for multiple comparisons using the Holm-Bonferroni method.

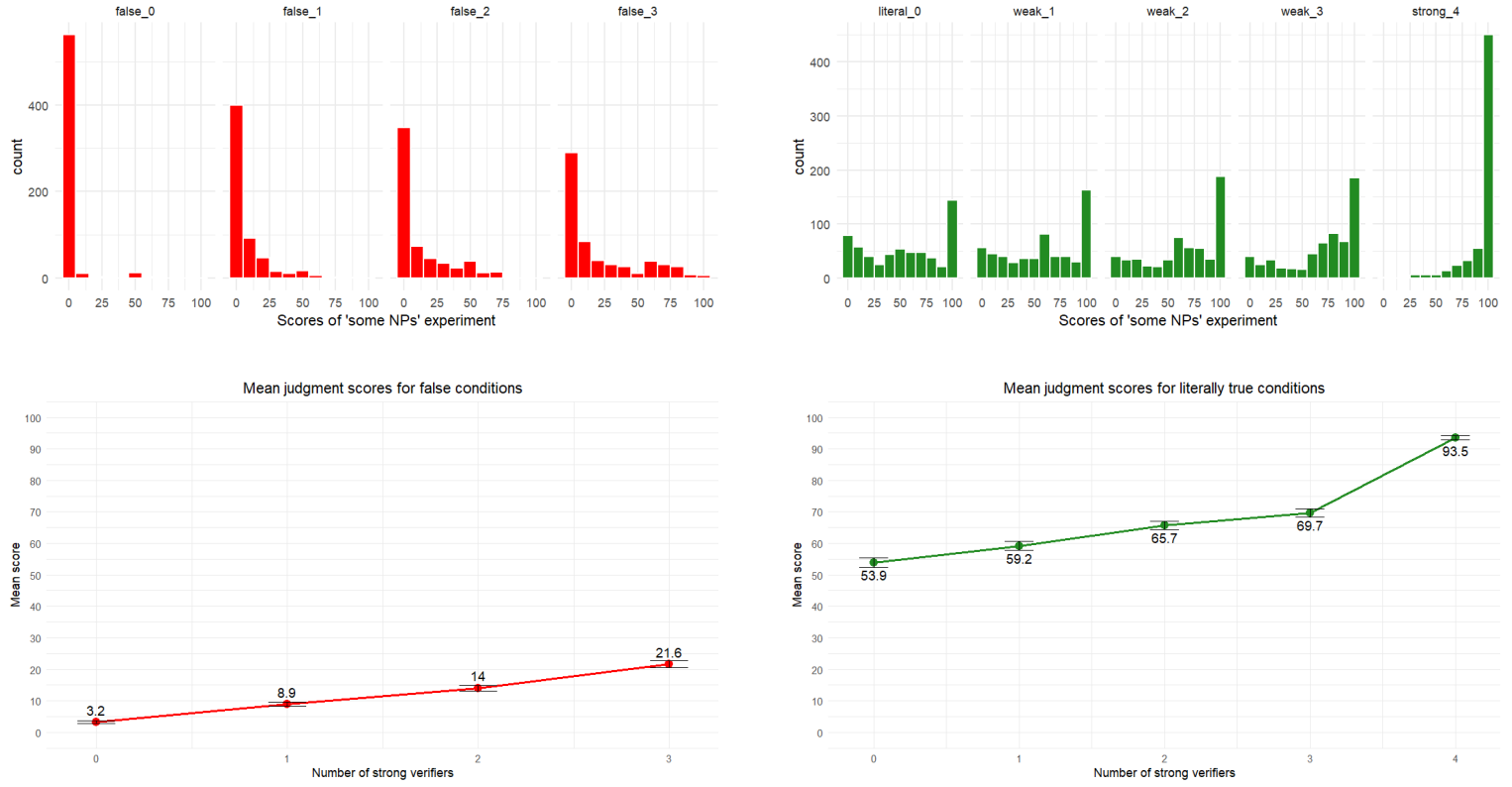


Figure 2.10: Results of the comprehension task with *some NPs* (continuous judgments)

The best-fitting model across all 9 conditions was the same as for the bare plurals experiment:

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

Table 2.7 reports statistical details of the best model.

Predictor	Estimate	Standard error	<i>t</i> -value
(Intercept)	3.361	1.284	2.618
c_{vrf}	5.716	0.272	20.989
c_{lit}	50.190	0.609	82.419
c_{str}	17.108	1.179	14.508

Table 2.7: Fixed effects of the best-fitting linear mixed-effects model (*some NPs* experiment, continuous judgments)

The second best-fitting model was, again:

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

with $\Delta BIC = 9$ between the best two models. This difference is considered strong evidence in favor of the better-ranked model.

The full model ranking using BIC is presented in Table 2.8 (for AIC results, see A.1 in Appendix A). The order of ranking using BIC is identical to the one obtained from the bare plurals data (Table 2.2). While the order differs between AIC and BIC (only by the relative positions of the models of rank 3 and 4), both criteria agree on the best two models.

Rank	Model formula	BIC
1	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{str}} + (1 \mid \text{participant})$	48869
2	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	48878
3	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + (1 \mid \text{participant})$	49067
4	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{weak}} + (1 \mid \text{participant})$	49070
5	$\text{response} \sim c_{\text{lit}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	49178
6	$\text{response} \sim c_{\text{lit}} + c_{\text{str}} + (1 \mid \text{participant})$	49283
7	$\text{response} \sim c_{\text{lit}} + c_{\text{weak}} + (1 \mid \text{participant})$	49899
8	$\text{response} \sim c_{\text{lit}} + (1 \mid \text{participant})$	50173
9	$\text{response} \sim c_{\text{vrf}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	50706
10	$\text{response} \sim c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	50707
11	$\text{response} \sim c_{\text{vrf}} + c_{\text{weak}} + (1 \mid \text{participant})$	51178
12	$\text{response} \sim c_{\text{weak}} + (1 \mid \text{participant})$	51250
13	$\text{response} \sim c_{\text{vrf}} + c_{\text{str}} + (1 \mid \text{participant})$	53206
14	$\text{response} \sim c_{\text{str}} + (1 \mid \text{participant})$	53385
15	$\text{response} \sim c_{\text{vrf}} + (1 \mid \text{participant})$	53761
16	$\text{response} \sim 1 + (1 \mid \text{participant})$	54771

Table 2.8: Model comparison by BIC (*some NPs* experiment, continuous judgments)

2.6.4 Discussion

Consistent with the findings from the cumulativity experiment (Section 2.4), baseline judgments for the **LITERAL-0** condition were lower with *some NPs* than with bare plurals. In response to our motivation for conducting this experiment, we still do not observe any qualitative difference between the **LITERAL-0** and **WEAK** conditions. As in the bare plurals experiment, the model comparison using BIC ranked the model without c_{weak} , but with all three other predictors, as the best. Therefore, it may not be the case that c_{weak} was left out from the best model because gradient concealed a shift between readings.

Furthermore, we conducted a post-hoc exploratory analysis consisting in model comparisons on the subset of data from the literally true conditions alone

(and consequently removing c_{lit} from the predictors). The motivation for this was the surprising observation that, in the subsequent study on *some NPs* with binary judgments, c_{weak} is included in the best model based on data subset to the literally true conditions. Comparison through BIC ranked as the best-fitting model:

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

The second best-ranked model included c_{weak} , with $\Delta BIC = 8$:

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

The next section will provide more discussion on this issue. We simply conclude that we have no evidence suggesting that the continuous nature of the response task favored the appearance of gradient effects to such an extent that they conceal the weak reading.

2.7 Comprehension study on *some NPs*: binary judgments

We conducted a second version of the experiment using binary truth-value judgments, in order to see whether gradient effects would still arise with binary judgments and to test whether the binary results could be predicted by a threshold-based model derived from the continuous judgments.

2.7.1 Methods

Methods were mostly identical to those of the comprehension study on *some NPs* with continuous judgments (Section 2.6), with the following differences:

1. **Participants.** After exclusions, we tested 200 adult participants recruited through Prolific (mean age 42.0; age range 18-76; 93 females). They had to meet the same eligibility criteria as in the previous experiments and must not have participated in the previous experiments.
2. **Stimuli.** The instructions accompanying each trial were: “Do you think the sentence below is true or false?”, with the options “false” on the left and “true” on the right.

2.7.2 Analyses

As the response type was binary, we analyzed the data using logistic mixed-effect models, and not linear mixed-effect models as we did with the continuous judgments. The rest of this analyses was identical to those conducted for *some NPs* with continuous judgments.

2.7.3 Results

Results are shown in Figure 2.11. There is hardly any observable gradience within one same level of reading (**FALSE** or **WEAK**), which justifies why some analyses conducted here need to be adapted. The adapted analyses will be further motivated and discussed in paragraph 2.7.4. This paragraph only presents results from preregistered analyses.

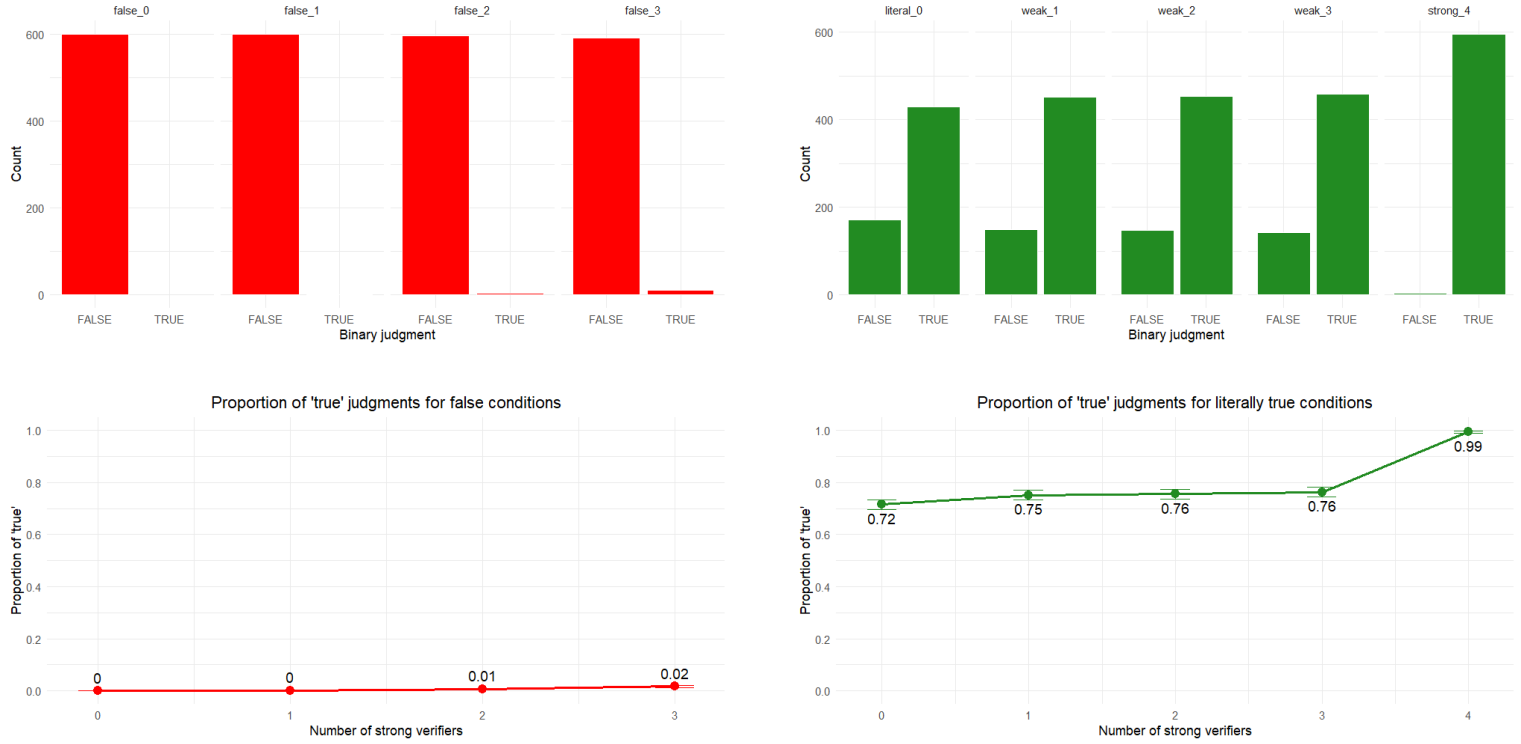


Figure 2.11: Results of the comprehension task with *some NPs* (binary judgments)

While fitting a mixed-effects model to the subset of literally true conditions, the random slope was removed due to failure of convergence. The model fit on literally true conditions (c_{vrf} as predictor) significantly outperformed the null model ($\chi^2(1) = 400.28, p < 10^{-15}$)⁷. As expected, the effect of c_{vrf} was positive ($b = 1.54, SE = 0.11, t = 14.03$).

However, within **WEAK** conditions alone, the LRT yielded $\chi^2(1) = 1.77$ and $p = 0.183$, suggesting that adding c_{vrf} did not significantly improve model fit compared to the null model.

⁷The p -values reported have been corrected for multiple comparisons using the Holm-Bonferroni method.

The best-fitting model across all 9 conditions was the same as for the previous experiments:

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

Table 2.9 reports statistical details of the best model.

Predictor	Estimate	Standard error	<i>t</i> -value
(Intercept)	-10.119	0.586	-17.279
c_{vrf}	0.505	0.097	5.213
c_{lit}	16.279	1.026	15.861
c_{str}	7.361	0.703	10.476

Table 2.9: Fixed effects of the best-fitting linear mixed-effects model (*some NPs* experiment, binary judgments)

The second best-fitting model was, again:

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

with $\Delta BIC = 8$ between the best two models. This difference is considered strong evidence in favor of the better-ranked model.

The full model rankings using BIC is presented in Table 2.10 (for AIC results, see Table A.2 in Appendix A). The order of ranking using BIC is not fully identical to either the one obtained through AIC, or the one obtained from the data of *some NPs* with continuous judgments (Table 2.8). However, the differences in ranking are not concerned with the best two models.

2.7.4 Discussion

Visually, three levels can be distinguished: the **FALSE** level, the **LITERAL+WEAK** level, and the **STRONG** level. Contrary to our experiments with continuous judgments, gradience is nearly non-existent here within the first two levels. Intuitively, this should indicate that the only relevant predictors are c_{lit} and c_{str} . However, model comparisons with BIC as well as AIC still select a model with the c_{vrf} factor as the best model.

The most plausible explanation is that the logistic model has limitations for probabilities near 0 or 1. In the **FALSE** conditions, the mean scores are very close to 0 and tiny differences in scores will be amplified in terms of log-odds. Because the log-odds difference between 0.01 and 0.02 is huge, the tiny ‘gradient’ effect in **FALSE** cases (which is theoretically irrelevant) could misleadingly drive the model to favor inclusion of the gradience predictor c_{vrf} . The equally

Rank	Model formula	BIC
1	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{str}} + (1 \mid \text{participant})$	1218
2	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	1226
3	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{weak}} + (1 \mid \text{participant})$	1227
4	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + (1 \mid \text{participant})$	1240
5	$\text{response} \sim c_{\text{lit}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	1433
6	$\text{response} \sim c_{\text{lit}} + c_{\text{str}} + (1 \mid \text{participant})$	1452
7	$\text{response} \sim c_{\text{lit}} + c_{\text{weak}} + (1 \mid \text{participant})$	1749
8	$\text{response} \sim c_{\text{lit}} + (1 \mid \text{participant})$	1820
9	$\text{response} \sim c_{\text{vrf}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	2959
10	$\text{response} \sim c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	3500
11	$\text{response} \sim c_{\text{vrf}} + c_{\text{weak}} + (1 \mid \text{participant})$	3819
12	$\text{response} \sim c_{\text{weak}} + (1 \mid \text{participant})$	3921
13	$\text{response} \sim c_{\text{str}} + (1 \mid \text{participant})$	5972
14	$\text{response} \sim c_{\text{vrf}} + c_{\text{str}} + (1 \mid \text{participant})$	5978
15	$\text{response} \sim c_{\text{vrf}} + (1 \mid \text{participant})$	6771
16	$\text{response} \sim 1 + (1 \mid \text{participant})$	7149

Table 2.10: Model comparison by BIC (*some NPs* experiment, binary judgments)

tiny gradient effect within **WEAK** conditions was shown through LRT to be non-significant.

As a sanity check, we conducted model comparisons only with the data from the literally true conditions. The predictor c_{lit} was therefore removed from the combinations of predictors. We expected the best model to include c_{str} , as the graph showed a stark increase in ratings. However, it was unclear, after visual inspection, whether c_{weak} was expected in the best model, given that the qualitative shift from **LITERAL** to **WEAK** was present, but small. The model ranking using BIC is presented in Table 2.11 (for AIC results, see A.3 in Appendix A).

The best-fitting model across the subset of 5 conditions was

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

Table 2.12 reports statistical details of the best model.

It is very interesting to observe that, among the many variants of experimental designs that we have presented, the weak reading only arose when the response task was binary. This could have an important theoretical consequence, as we might have uncovered an argument in favor of the weak reading being accessed, contra the best models yielded by all of the continuous response tasks. However, more careful investigation would be necessary to understand the potential link between the detection of the weak reading and the nature of the task.

Rank	Model formula	BIC
1	$\text{response} \sim c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	937
2	$\text{response} \sim c_{\text{vrf}} + c_{\text{str}} + (1 \mid \text{participant})$	941
3	$\text{response} \sim c_{\text{vrf}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	944
4	$\text{response} \sim c_{\text{str}} + (1 \mid \text{participant})$	954
5	$\text{response} \sim c_{\text{vrf}} + c_{\text{weak}} + (1 \mid \text{participant})$	1170
6	$\text{response} \sim c_{\text{vrf}} + (1 \mid \text{participant})$	1187
7	$\text{response} \sim c_{\text{weak}} + (1 \mid \text{participant})$	1497
8	$\text{response} \sim 1 + (1 \mid \text{participant})$	1579

Table 2.11: Model comparison by BIC (*some NPs* experiment, binary judgments, subset of literally true conditions)

Predictor	Estimate	Standard error	<i>z</i> -value	<i>p</i> -value
(Intercept)	8.606	0.664	12.961	$< 2e^{-16}$
c_{weak}	1.394	0.289	4.816	$1.46e^{-6}$
c_{str}	9.426	0.738	12.770	$< 2e^{-16}$

Table 2.12: Fixed effects of the best-fitting logistic mixed-effects model (*some NPs* experiment, binary judgments, literally true conditions)

To explore the relationship between the continuous and binary versions of the experiment, we sought to identify whether the actual binary responses could be predicted by applying a binarization threshold to the continuous responses. Conceptually, this would mean that participants have an underlying continuous judgment to which they apply a threshold when asked for a binary response. To simulate this, we applied a manually selected threshold to the continuous scores, binarizing each participant’s responses individually before averaging them. We then computed the Pearson correlation coefficient between these simulated binary values and the experimental binary values, as a measure of how well the binarized continuous data aligns with the original binary data. The correlation score provides an indication of how consistent the two response formats are. For simplicity, we only tried with binarization thresholds by increments of 10. The results are presented in Table 2.13:

Threshold	90	80	70	60	50	40	30	20	10
Pearson <i>r</i>	0.902	0.923	0.943	0.965	0.980	0.984	0.982	0.973	0.954

Table 2.13: Pearson correlation coefficient at varying thresholds.

The highest correlation was obtained using a threshold of 40 (on a scale of 1-100). However, all tested thresholds yielded very high correlation scores, exceeding 0.9. To better understand this result, we examined the means of the simulated binary data in each condition for each threshold. We observed that higher thresholds tended to produce good predictions in the **FALSE** conditions, but deviated from the actual means in literally true conditions. Conversely for low thresholds. Therefore, the high overall correlations scores can obscure important variations in predictive accuracy across conditions. This shows the limitations of relying only on one threshold and on a correlation score as an evaluation metric.

We then performed a variation on this exploratory analysis to simulate ‘ternarization’ using two thresholds, similar to the framework of trivalent theories. We used varying pairs of low and high thresholds. For simplicity, all thresholds were multiples of 10 taken between 10 and 90. For each (low, high) pair such that the gap between thresholds was at least 10, we mapped continuous cursor responses to binary predictions: values above the high threshold were interpreted as “true”, those below the low threshold as “false”, and values in between were resolved by a random choice between true and false. The table (2.14) shows the top ten threshold pairs in terms of correlation.

Threshold pair	(30,40)	(30,50)	(40,50)	(20,40)	(40,60)	(30,60)	(10,40)	(10,50)	(20,50)	(20,30)
Pearson r	0.982	0.982	0.981	0.981	0.979	0.979	0.978	0.978	0.977	0.977

Table 2.14: Top ten Pearson correlation coefficients by pairs of thresholds.

As we can see, many pairs of thresholds closely correlate with the actual binary data and we cannot draw a definitive conclusion about the best pair of thresholds. All things considered, a threshold-based model is likely insufficient to predict binary judgments from continuous ones.

Chapter 3

Comparison with Mandarin

This chapter introduces a cross-linguistic perspective by examining Mandarin Chinese, a language in which number marking is optional, unlike in English. In Mandarin, bare nouns are widely used alongside singular and plural forms. A more thorough description of plural definites (notably, the plural suffix *-men*) and indefinites in Mandarin has been made in Rong 2024. This chapter focuses on the plural classifier *xie*.

We conducted a Mandarin version of the previous experiments involving plural indefinites in English. The goal was to detect potential differences in plural comprehension between the two languages. Such differences may shed light on how competition between alternative forms operates in languages with different number marking systems.

3.1 A language with optional number marking

Mandarin has a threefold number expression system:

- Bare nouns (BNs), which are number-neutral when used as indefinites (Zhang 2014; Cheng and Sybesma 1999, a.o.).
- [one + CL + NP], the form for singular indefinites¹ (e.g., yi-gè 一个). We use CL as the generic abbreviation for atomic classifiers. [one + CL + NP] triggers a uniqueness inference in UE environments.
- [one + *xie* + NP], a form for plural indefinites². [one + *xie*] is the gloss for yi-xiē 一些, where xiē (些) is the plural classifier. [one + *xie* + NP] triggers a multiplicity inference in UE environments.

¹If *one* is replaced with another number *N*, then [*N* + CL + NP] expresses a plural indefinite, but we are not concerned with this form in our analysis.

²Note that the suffix *-men* 们 marks the NP as a definite plural if the NP is not preceded by *xie*. *-men* is neither a necessary nor a sufficient marker of a plural indefinite, despite being a well-known ‘plural marker’ in Mandarin. For more discussions on *-men*, see Rong 2024.

In production, bare nouns are widely preferred, as they are under-specified for number.

When *xie* is used, it takes the place of the atomic classifier usually associated with the noun. The only number that can precede *xie* is 一 *yī* (*one*), even though it does not contribute any meaning of uniqueness. In formal Mandarin, *yī* should always be written, but it is often omitted orally regardless of what classifier comes right after (and it is the only number that may be omitted). We will be writing the formal versions of all the glossed sentences.

When embedded in a DE environment³, [one + *xie* + NP] no longer triggers a multiplicity inference:

- (48) 每 当 小 明 看 到 一 些 兔 子,
měi dāng xiǎo-míng kàn dào yī xiē tù-zi
each when Xiao-míng see CMPL one ~~XIE~~ rabbit
他 都 会 高 兴。
tā dōu huì gāo-xìng
he DOU will happy

‘Each time Xiao-ming see some rabbits, he’s happy.’

Intuitively, (48) suggests that Xiao-ming is happy as soon as he sees at least one rabbit⁴. If we treat [one + *xie* + NP] as weak plural, we can expect it to exhibit the same range of possible readings as English bare plurals under universal quantification: literal, weak, strong. However, we also have the intuition that the sentence glossed as “every box contains [one + *xie* + NP]” is judged true only in situations that support the strong reading. In this study, we empirically test whether this intuition is shared by participants, or whether additional readings are available to them.

3.2 Comprehension study on *xie*: continuous judgments

3.2.1 Methods

Participants

We initially aimed at testing 200 participants *after* exclusions. The sample size was based on power estimates using the data collected through the English

³The DE environment must be other than negation, as [one + *xie* + NP] is a Positive Polarity Item (PPI). The account from Ahn, Saha, and Sauerland 2020 predicts that plurals should be PPIs in languages with optional number marking, as is the case in Mandarin.

⁴The majority of native informants agree with our introspective judgment, but some speakers will disagree. It seems that there is less consensus on [one + *xie* + NP] than on *some* NPs.

bare plurals experiment. Following our bootstrap analysis, a sample size of 200 yielded a statistical power over 80% in order to have a difference in BIC greater than 4 between the best two models, and a sample size of 150 yielded a similar statistical power for a difference in BIC greater than 2.

However, recruitment proved to be more difficult than anticipated. Part of our participants had first been recruited through direct contact and snowball sampling (based on voluntary participation). Only 23 participants completed the experiment with this first recruitment method. We decided to recruit the remaining participants through Prolific (mean age 30.6; age range 18-70; 102 females). They were paid £0.60 for their participation. We still did not reach our target of 200 after exclusions, as only 209 participants in total completed the experiment and there remained less than 200 after exclusions.

Prolific participants had to fulfill the eligibility criteria of reporting Chinese to be their ‘primary language’ and ‘earliest language in life’.

Procedure and stimuli

The stimuli are identical to those of the English bare plurals experiment (Section 2.3), and the instructions are the exact Chinese translation of the English version. Here is a gloss of the Chinese sentence in the trials:

- (49) 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].’

3.2.2 Analyses

Participants who completed the full experiment were excluded if:

1. they answered “no” to the preliminary question of the experiment (translated into Chinese) “Is Chinese your native language?” (16 participants);
2. they failed the attention check (31 participants);
3. they did not rate every trial of the condition **STRONG-4** higher than every trial of the condition **FALSE-0** (7 participants).

After exclusions, 155 participants were included in the analyses. Despite difficulties with recruitment, this sample size was still satisfactory.

We conducted the same analyses as for the English bare plurals experiment.

3.2.3 Results

Results are shown in Figure 3.1.

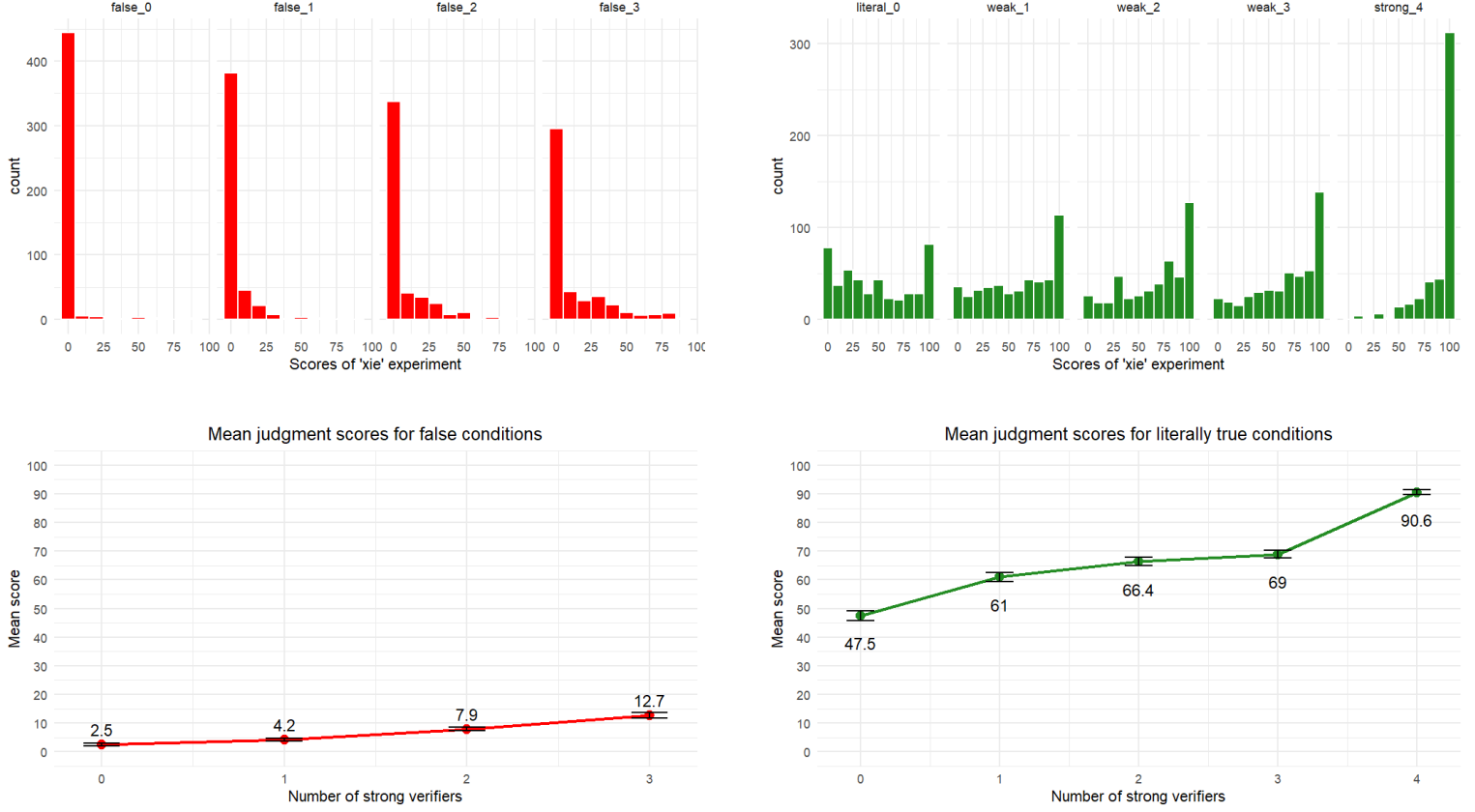


Figure 3.1: Results of the comprehension task with *xie* (continuous judgments)

While fitting a mixed-effects model to the subset of literally true conditions, the random slope was removed due to failure of convergence. The model fit on literally true conditions (c_{vrf} as predictor) significantly outperformed the null model ($\chi^2(1) = 858.13, p < 10^{-15}$)⁵. As expected, the effect of c_{vrf} was positive ($b = 9.41, SE = 0.29, t = 32.44$).

A gradient effect was also found within **WEAK** conditions alone ($\chi^2(1) = 38.34, p < 10^{-9}$), where the random slope allowed model convergence and was kept.

The best-fitting model across all 9 conditions was different from the best model fitting the data of all our experiments with English plural indefinites:

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

⁵The p -values reported have been corrected for multiple comparisons using the Holm-Bonferroni method.

Table 3.1 reports statistical details of the best model.

Predictor	Estimate	Standard error	<i>t</i> -value
(Intercept)	1.4727	1.3463	1.094
c_{vrf}	3.5842	0.3517	10.191
c_{lit}	46.0392	1.1665	39.468
c_{weak}	10.7614	1.2843	8.379
c_{str}	17.9750	1.2843	13.996

Table 3.1: Fixed effects of the best-fitting linear mixed-effects model
(*xie* experiment, continuous judgments)

The second best-fitting model was:

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

with $\Delta BIC = 62$ between the best two models. This difference is considered very strong evidence in favor of the better-ranked model.

The full model ranking using BIC and AIC is presented in Table 3.2. The order of ranking was identical between the two criteria.

3.2.4 Discussion

Visually, the graphs revealed clear gradient effects, and most importantly, a level of weak reading that was absent in our data from English plural indefinites. The best-fitting model according to BIC (and AIC) includes all four predictors, which raises the question: why is the weak reading detected in Mandarin but not in English? Taken at face value, this could suggest that [one + *xie*] in Mandarin elicits more fine-grained levels of interpretation than English *some NPs*. Two factors may help explain this difference:

1. In a three-way number marking system, competition with the bare noun may lead classifier-marked singulars and plurals to be used in more specific contexts.
2. The participant samples for two languages are likely from different backgrounds. The Chinese-speaking sample may represent a more highly educated population compared to the English-speaking sample, as the vast majority of Chinese participants were not residing in China but were studying or working abroad.

Rank	Model formula	BIC	AIC
1	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	37453	37409
2	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{str}} + (1 \mid \text{participant})$	37515	37477
3	$\text{response} \sim c_{\text{lit}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	37548	37509
4	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{weak}} + (1 \mid \text{participant})$	37636	37598
5	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + (1 \mid \text{participant})$	37670	37639
6	$\text{response} \sim c_{\text{lit}} + c_{\text{str}} + (1 \mid \text{participant})$	37802	37770
7	$\text{response} \sim c_{\text{lit}} + c_{\text{weak}} + (1 \mid \text{participant})$	38040	38009
8	$\text{response} \sim c_{\text{lit}} + (1 \mid \text{participant})$	38473	38448
9	$\text{response} \sim c_{\text{vrf}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	38762	38724
10	$\text{response} \sim c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	38807	38775
11	$\text{response} \sim c_{\text{vrf}} + c_{\text{weak}} + (1 \mid \text{participant})$	39166	39134
12	$\text{response} \sim c_{\text{weak}} + (1 \mid \text{participant})$	39170	39145
13	$\text{response} \sim c_{\text{vrf}} + c_{\text{str}} + (1 \mid \text{participant})$	41433	41401
14	$\text{response} \sim c_{\text{str}} + (1 \mid \text{participant})$	41539	41513
15	$\text{response} \sim c_{\text{vrf}} + (1 \mid \text{participant})$	41866	41841
16	$\text{response} \sim 1 + (1 \mid \text{participant})$	42572	42553

Table 3.2: Model comparison by BIC and AIC (*xie* experiment, continuous judgments)

Synthesis and closing remarks

Summary and methodological discussions

We repeat here the three core questions that guided our inquiries, along with the answers drawn from our experimental results:

Core Theoretical Question 1

What are the available readings?

Core Methodological Question

Experimentally, how can we disentangle readings from gradient effects?

All things considered, our results do not provide definitive evidence for either the existence or the non-existence of the weak reading. Based on data from English comprehension, model comparisons showed that the factor c_{weak} coding for the weak reading does not improve the fit of the model. The best-fitting model for our data on bare plurals and *some NPs* was the following, in the case of continuous judgments:

$$(50) \quad \text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{str}} + (1 \mid \text{participant})$$

This seems to go in favor of theories that do not predict the weak reading, namely Zweig’s approach and the two trivalent approaches. If this is the case, further investigations are still necessary to adjudicate between those approaches. One possible method would be to compare them based on their predictions on context-sensitivity of bare plurals depending on the polarity of the sentence (positive or negative). This method was used by Jiang and Sudo 2023 for different purposes.

An important caveat here is that the weak reading was in fact detected in literally true conditions in the binary *some NPs* experiment. This underlines the methodological difficulties in designing a truly conclusive experiment to detect available readings.

Another methodological challenge was how to incorporate gradience into our models. We have translated gradience into discrete values of the factor c_{vrf} , which simply coded for the number of strong verifiers (i.e. boxes with several objects). Counting strong verifier boxes seemed like the most straightforward way to incorporate gradience, but alternative representations of gradience could also have been considered. For instance, we could have adopted a model similar to that proposed by Chemla and Spector 2014, where the w_i 's are the weights of different readings and α, β, γ are the weights of different factors. Adapted with our notations and with the assumption that the weak reading does not improve the model, this alternative model would be expressed as:

$$\begin{aligned}
 \text{response} \sim & w_1 \times (\alpha \times \text{Truth-value of the literal reading} \\
 & + \beta \times \text{Distance to the closest true case for the literal reading} \\
 & + \gamma \times \text{Distance to the closest false case for the literal reading}) \\
 (51) \quad & w_2 \times (\alpha \times \text{Truth-value of the strong reading} \\
 & + \beta \times \text{Distance to the closest true case for the strong reading} \\
 & + \gamma \times \text{Distance to the closest false case for the strong reading})
 \end{aligned}$$

It is not straightforward to see if our best model's formula (50) is equivalent to an expression of the form (51).

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?

Gradient effects are indeed also observed in Mandarin, contrary to our initial intuition that only the strong reading would be accessible. The main difference with the English data is that the best-fitting model includes c_{weak} along with all other predictors. The mechanisms of plural interpretation likely depend on the alternative forms available in every given number marking system. Further theoretical work may help determine whether the availability of the weak reading in Mandarin is linked to the existence of bare nouns in the number marking system and the fact that bare nouns are, at least intuitively, widely preferred in production.

Conclusion

Given the goals of this thesis, our core questions have been answered, but also raise many questions for future research. Among the results that require more explanation and refinement, there are:

- **The source of gradient effects:** As we have mentioned, Chemla and Spector 2014 model typicality as a weighted sum of readings, combining truth-values with distance to the nearest scenario that makes a certain reading true or false. However, the very notion of distance between two scenarios raises conceptual challenges. Even in the simple case of a sentence with no ambiguity and a single clear true reading (e.g., “Every box has [several NPs]”), there are multiple ways to interpret distance to the scenario making the sentence true. For instance, it might be measured by the number of boxes that must be modified to make the strong reading true (analogous to Hamming distance), or by the total number of geometric shapes that need to be added. One might interpret distance to a given scenario as reflecting degrees of *verisimilitude*, for which several formal measures exist (see e.g. Süskind 2024 for a recent overview). However, the situation becomes even more complex when ambiguity arises between multiple readings, since partial truth and ambiguous truth are fundamentally different phenomena that call for separate analyses (Süskind 2024, Appendix B). Our results show the complexity of analyzing data that reflect both partial truth (within a given reading, how close we are to a certain situation) and ambiguous truth (which readings are accessed).

There are also alternative explanations for gradience. For instance, van Tiel and Geurts 2014 model typicality via individual exemplars: a scenario is a good fit for “Each box contains books” when each box contains a typical number of “books” (i.e., more than one). Testing this hypothesis against the one of Chemla and Spector 2014 would require an independent measure of the typicality of bare plurals as a function of the number of individuals.

- **Acceptability of the cumulative reading:** we observed, in Section 2.4, different levels of acceptability for cumulative readings depending on the type of plural expression (bare plurals, *some NPs*, *several NPs*). While this puzzle did not impact the analysis of our other results, it warrants an independent explanation.
- **Linking continuous and binary responses:** as we have seen in Paragraph 2.7.4, one cannot yet conclude whether binary responses arise from an underlying continuous judgment which would then be compared to a certain threshold. We have tried simple simulations of binarizing and ‘ternarizing’ our actual data from continuous judgments, and we compared it to the actual data from binary judgments. This exploratory analysis was somewhat rough, in the sense that we assumed that all participants have the same underlying threshold. A more fine-grained analysis would consist in finding a model predicting the binary responses from the continuous responses.
- **Linking language production and language comprehension:** we have only conducted one production study at the very beginning of this project (Section 2.2), and we have focused on comprehension studies afterwards. Whether

speakers behave in a Bayesian way across production and comprehension remains a broader and interesting question, and its study would be crucial to a cognitive model of language use.

Further empirical and theoretical work can extend this research in the following directions:

- **Other lexical scales:** we can revisit other lexical items which are considered to give rise to scalar implicatures, given that the distribution of multiplicity inferences should mirror the distribution of scalar implicatures, according to implicature-based approaches. Chemla and Spector 2011 found gradient effects in their experimental results aiming at providing evidence for the existence of local scalar implicatures. They argued that these gradient effects did not threaten the conclusion that strong readings exist (whereby “Every student read some of the books” means that every student read some but not all of the books), but did not, in fact, consider the weak reading from this perspective.
- **Beyond universals:** as we expand the empirical studies to other lexical scales, we can also explore additional syntactic environments, in particular non-monotonic ones (e.g. the scope of *exactly N*) that are crucial for detecting local scalar implicatures.
- **Role of the Question Under Discussion (QUD):** all theories predict that context, in particular the QUD, modulates the readings of plural expressions, but differ in terms of their precise predictions (Jiang and Sudo 2023). Therefore, the role of QUDs needs to be investigated more systematically.
- **Probabilistic modeling:** our experiments placed plural interpretation in contexts without communication, and one may wonder how results might differ in interactive, communicative settings. To this end, we can refine our models in frameworks like Rational Speech Act theory (Cremers, Wilcox, and Spector 2023), which treats communication as recursive speaker-listener probabilistic reasoning.

Bibliography

- Ahn, Dorothy, Ankana Saha, and Uli Sauerland (2020). “Positively polar plurals: Theory and predictions”. In: *Semantics and linguistic theory*, pp. 450–463.
- Bassi, Itai, Guillermo Del Pinal, and Uli Sauerland (2021). “Presuppositional exhaustification”. Unpublished manuscript.
- Chemla, Emmanuel and Benjamin Spector (2011). “Experimental evidence for embedded scalar implicatures”. In: *Journal of semantics* 28.3, pp. 359–400.
- (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.
- Chierchia, Gennaro (2006). “Broaden your views: Implicatures of domain widening and the “logicality” of language”. In: *Linguistic inquiry* 37.4, pp. 535–590.
- Cremers, Alexandre, Ethan G Wilcox, and Benjamin Spector (2023). “Exhaustivity and Anti-Exhaustivity in the RSA Framework: Testing the Effect of Prior Beliefs”. In: *Cognitive Science* 47.5, e13286.
- Doron, Omri and Jad Wehbe (2022). “A constraint on presupposition accommodation”. In: *Proceedings of the 23rd Amsterdam Colloquium*. Vol. 23.
- 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.

- Jiang, Yizhen and Yasutada Sudo (2023). “Putting bare plurals into context”. In: *Hnm2—gaps and imprecision in natural language semantics: homogeneity effects and beyond*.
- Križ, Manuel (2017). “Bare plurals, multiplicity, and homogeneity”. Unpublished manuscript.
- Križ, Manuel and Benjamin Spector (2021). “Interpreting plural predication: Homogeneity and non-maximality”. In: *Linguistics and Philosophy* 44, pp. 1131–1178.
- Landman, Fred (2000). “Plural roles, scope and event types”. In: *Events and Plurality: The Jerusalem Lectures*, pp. 177–221.
- Link, Godehard et al. (1983). “The logical analysis of plurals and mass terms: A lattice-theoretical approach”. In: *Formal semantics: The essential readings* 127, p. 147.
- Link, Godehard (1990). “Algebraic semantics in language and philosophy”. In: *CSLI Lecture Notes* 74.
- Raftery, Adrian E (1995). “Bayesian model selection in social research”. In: *Sociological methodology*, pp. 111–163.
- Rong, Claire (2024). “Plurality in Mandarin Chinese”. Supervisors: Benjamin Spector and Chang Liu. M1 thesis in Chinese Studies. École Normale Supérieure de Lyon.
- Sauerland, Uli (2003). “A new semantics for number”. In: *Semantics and linguistic theory*.
- 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.
- Süskind, Jakob (2024). “On the theory of verisimilitude”. PhD thesis. École Normale Supérieure.

- 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.
- (2008). “Dependent plurals and plural meaning”. PhD thesis. New York University.

Appendix A

AIC tables

Rank	Model formula	AIC
1	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{str}} + (1 \mid \text{participant})$	48830
2	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	48831
3	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{weak}} + (1 \mid \text{participant})$	49031
4	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + (1 \mid \text{participant})$	49034
5	$\text{response} \sim c_{\text{lit}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	49139
6	$\text{response} \sim c_{\text{lit}} + c_{\text{str}} + (1 \mid \text{participant})$	49251
7	$\text{response} \sim c_{\text{lit}} + c_{\text{weak}} + (1 \mid \text{participant})$	49866
8	$\text{response} \sim c_{\text{lit}} + (1 \mid \text{participant})$	50147
9	$\text{response} \sim c_{\text{vrf}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	50666
10	$\text{response} \sim c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	50674
11	$\text{response} \sim c_{\text{vrf}} + c_{\text{weak}} + (1 \mid \text{participant})$	51145
12	$\text{response} \sim c_{\text{weak}} + (1 \mid \text{participant})$	51224
13	$\text{response} \sim c_{\text{vrf}} + c_{\text{str}} + (1 \mid \text{participant})$	53173
14	$\text{response} \sim c_{\text{str}} + (1 \mid \text{participant})$	53359
15	$\text{response} \sim c_{\text{vrf}} + (1 \mid \text{participant})$	53734
16	$\text{response} \sim 1 + (1 \mid \text{participant})$	54751

Table A.1: Model comparison by AIC (*some NPs* experiment, continuous judgments)

Rank	Model formula	AIC
1	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{str}} + (1 \mid \text{participant})$	1185
2	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	1186
3	$\text{response} \sim c_{\text{lit}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	1194
4	$\text{response} \sim c_{\text{lit}} + c_{\text{str}} + (1 \mid \text{participant})$	1213
5	$\text{response} \sim c_{\text{vrf}} + c_{\text{lit}} + c_{\text{weak}} + (1 \mid \text{participant})$	1400
6	$\text{response} \sim c_{\text{lit}} + c_{\text{vrf}} + (1 \mid \text{participant})$	1426
7	$\text{response} \sim c_{\text{lit}} + c_{\text{weak}} + (1 \mid \text{participant})$	1723
8	$\text{response} \sim c_{\text{lit}} + (1 \mid \text{participant})$	1800
9	$\text{response} \sim c_{\text{vrf}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	2926
10	$\text{response} \sim c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	3474
11	$\text{response} \sim c_{\text{vrf}} + c_{\text{weak}} + (1 \mid \text{participant})$	3792
12	$\text{response} \sim c_{\text{weak}} + (1 \mid \text{participant})$	3901
13	$\text{response} \sim c_{\text{vrf}} + c_{\text{str}} + (1 \mid \text{participant})$	5951
14	$\text{response} \sim c_{\text{str}} + (1 \mid \text{participant})$	5952
15	$\text{response} \sim c_{\text{vrf}} + (1 \mid \text{participant})$	6752
16	$\text{response} \sim 1 + (1 \mid \text{participant})$	7136

Table A.2: Model comparison by AIC (*some NPs* experiment, binary judgments)

Rank	Model formula	AIC
1	$\text{response} \sim c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	913
2	$\text{response} \sim c_{\text{vrf}} + c_{\text{weak}} + c_{\text{str}} + (1 \mid \text{participant})$	914
3	$\text{response} \sim c_{\text{vrf}} + c_{\text{str}} + (1 \mid \text{participant})$	917
4	$\text{response} \sim c_{\text{str}} + (1 \mid \text{participant})$	936
5	$\text{response} \sim c_{\text{vrf}} + c_{\text{weak}} + (1 \mid \text{participant})$	1146
6	$\text{response} \sim c_{\text{vrf}} + (1 \mid \text{participant})$	1169
7	$\text{response} \sim c_{\text{weak}} + (1 \mid \text{participant})$	1479
8	$\text{response} \sim 1 + (1 \mid \text{participant})$	1567

Table A.3: Model comparison by AIC (*some NPs* experiment, binary judgments, subset of literally true conditions)

Appendix B

Preregistrations

In order:

1. Production experiment in English
2. Comprehension experiment in English: bare plurals, continuous judgments
3. Comprehension experiment in English: cumulativity of bare plurals, *some NPs* and *several NPs*, continuous judgments
4. Comprehension experiment in English: *several NPs*, continuous judgments
5. Comprehension experiment in English: *some NPs*, continuous judgments
6. Comprehension experiment in English: *some NPs*, binary judgments
7. Comprehension experiment in Mandarin: *xie*, continuous judgments

Number marking in universally quantified statements: a production study.

Benjamin Spector, Claire Rong

1. Study design

1.1 Participants

We will test adult participants who report English to be their ‘first language’, ‘primary language’ and ‘earliest language in life’.

1.2 Stimuli

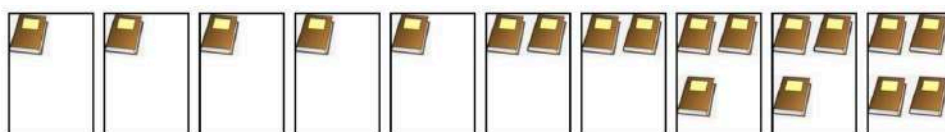
Each participant will be shown a single picture containing 10 boxes. Each box will contain between 1 and 4 objects, with the number of objects possibly differing between boxes. There will be 11 possible object types, all of which will be familiar to participants. Across all possible stimuli, the number of boxes containing a unique object will vary from 0 to 10. The experimental condition is defined as the number of boxes containing a unique object.

Each condition is instantiated 11 times, with 11 different objects (apples, bikes, birds, books, chairs, flowers, houses, pencils, rabbits, stars, trees). The stimuli are available in the stimuli folder (in this OSF project).

1.3 Procedure

Each participant completes exactly one trial. They are asked to complete a sentence of the form ‘Every box contains....’, based on the picture they see.

Here is an example of a trial:



Please observe the picture and complete the sentence below.

Every box contains

Next

On the second page (from which participants cannot navigate back), they answer the following question: *How many Xs were there in the box(es) with the fewest Xs?* where "X" is replaced by the name of the object shown.

The third and final page includes an attention check.

2. Sampling plan

2.1 Data collection procedure

Participants will be recruited through the online platform Prolific and paid £0.3 for their participation.

2.2 Inclusion and exclusion criteria

Participants will be excluded from the analyses if:

- i) they do not complete the full experiment;
- ii) they fail the attention check;
- iii) they give a wrong answer to the second question (*How many Xs were there in the box(es) with the fewest Xs?*).

2.3 Sample size

We estimate that a sample size of 250 participants is appropriate. This will allow us to collect between 15 and 20 valid data points for each of the 11 conditions. No further justification for the sample size is provided, as we have no prior estimate of the potential effect size.

3. Hypotheses

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 expect a gradient effect, such that the proportion of boxes containing a unique object will positively correlate with the proportion of singular nouns used and negatively correlate with the proportion of bare plurals used.

4. Statistical tests

We will subset the answers for which either a simple singular NP was used ("a NP", e.g., 'a rabbit', 'a white rabbit', not including numerals and modified numerals such as 'one' or 'at least one', etc.) or a bare plural NP was used (NPs, e.g. 'rabbits', 'white rabbits').

Relative to this restricted data set, we will run two tests:

- 1) We will fit a logistic model (using maximum likelihood) to predict the log-odds of choosing a singular noun as a function of the number of boxes that contain exactly one object (a factor coded as numeric).

Model: $\log(p(\text{choosing singular})/p(\text{choosing plural})) = a + b.n$, where n is the number of boxes with only one object (a factor coded as numeric).

We will compare this model with a null model by means of a likelihood ratio test (LRT).

Null Model: $\log(p(\text{choosing singular})/p(\text{choosing plural})) = a'$

We predict b to be positive and the LRT to return a significant p -value.

- 2) To make sure that the effect we found is not only due to the most extreme conditions (0 box with just one object, or all boxes with just one object), we will run the same test on the data without these extreme conditions.

We will correct for multiple comparisons by means of the Holm Bonferroni method

We will also run a number of exploratory analyses in order to have a fine-grained understanding of whatever gradient effects are found.

Bare plurals in universally quantified statements: a comprehension study.

Benjamin Spector, Claire Rong

Due to miscommunication between the two collaborators, data was collected before this preregistration was posted, and preliminary descriptive analyses were conducted. In this sense, the statistical analyses we report here are post hoc, even though they were in fact planned in advance.

1. Background

There exist several approaches to the semantic and pragmatic interpretation of plural indefinites: the implicature approach (Sauerland 2003, Spector 2007, Zweig 2007), which comes in many different variety, the homogeneity approach (Kriz 2017), the presuppositional approach (a version of which is also an implicature-based approach, Bassi, del Pinal & Sauerland 2021). These accounts all agree that the multiplicity inference triggered by plural indefinites is not a standard entailment. They all predict that the multiplicity inference is not part of the content that is negated in a simple negative sentence. Theories differ from each other (even when they belong to the same class of approaches) with respect to cases where a plural indefinite is under the scope of a universal quantifier. We have run some exploratory experiments and in production we observed gradient effects regarding the choice between sentences like *Every box contains pencils* vs *Every box contains a pencil*, depending on the proportion of boxes that contain one or several pencils (cf. preregistration <https://osf.io/5dxe7>). We want to investigate gradience in comprehension - a gradient effect whereby the proportion of boxes with several pencils (in this case) influence behavior in a truth-value judgment task is not predicted by existing theories (except maybe Enguehard 2024).

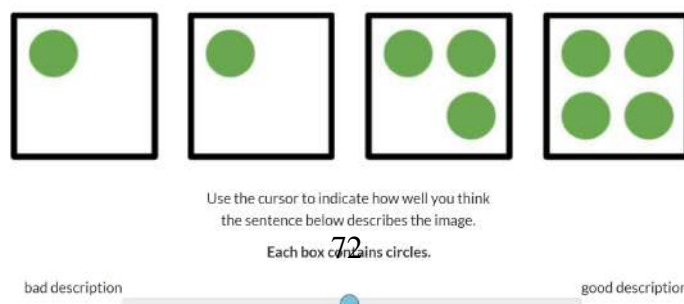
2. Study design

2.1 Participants

We will test adult participants who report English to be their ‘first language’, ‘primary language’ and ‘earliest language in life’.

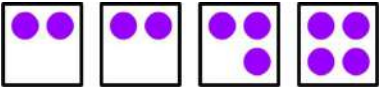
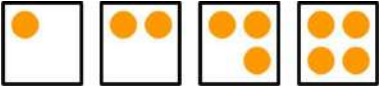
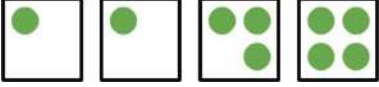
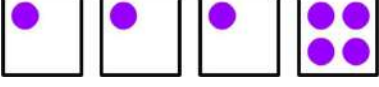





2.2 Stimuli and procedure

A trial consists of a picture and a sentence: the sentence follows a fixed structure “Each box contains [bare plural]” and is paired with a picture made up of 4 boxes containing between 0 and 4 geometric shapes. Here is an example of a trial:



There are two types of conditions in the experiment:

1. Literal truth conditions
 - A sentence is literally false if at least one box is empty (4 conditions).
 - A sentence is literally true otherwise (5 conditions).
2. Experimental conditions
 - A box containing multiple shapes is called a strong verifier.
 - Conditions are labeled based on the strongest reading they satisfy (FALSE, LITERAL, WEAK, or STRONG), followed by the number of strong verifiers in the image.

stimuli pictures (examples with circles)	label for condition	readings
	STRONG-4	true for STRONG+WEAK+LITERAL readings
	WEAK-3	true for WEAK+LITERAL readings
	WEAK-2	true for WEAK+LITERAL readings
	WEAK-1	true for WEAK+LITERAL readings
	LITERAL-0	true for LITERAL reading
	FALSE-3	false for LITERAL reading
	FALSE-2	false for LITERAL reading
	FALSE-1	false for LITERAL reading
	FALSE-0	false for LITERAL reading

Each trial is accompanied by the same instruction: “Use the cursor to indicate how well you think the sentence below describes the image”. In the sentence “Each box contains [bare plural]”, the [bare plural] was one of the three following words, always corresponding to the geometric shape shown in the picture: *circles*, *triangles*, or *squares*. The cursor moves on a continuous scale ranging from “bad description” (left extremity) to “good description” (right extremity). Participants’ cursor ratings are saved as integers between 1 and 100, but the rating number is not visible to participants themselves.

Each participant completes 27 randomized trials, with each condition appearing three times, using three different geometric shapes in three different colors (coupling of shapes/colors is reported in the table below). The stimuli are available in the stimuli folder (in this OSF project).

	green circles	orange circles	purple circles	green squares	orange squares	purple squares	green triangles	orange triangles	purple triangles
FALSE-1		x		x					x
FALSE-2			x		x		x		
FALSE-3	x					x		x	
LITERAL-0		x		x					x
WEAK-1			x		x		x		
WEAK-2	x					x		x	
WEAK-3		x		x					x
STRONG-4			x		x		x		

The final page contains an attention check.

3. Sampling plan

3.1 Data collection procedure

Participants will be recruited through the online platform Prolific and paid £0.75 for their participation.

3.2 Inclusion and exclusion criteria

Participants will be excluded from the analyses if:

- i) they do not complete the full experiment;
- ii) they fail the attention check;
- iii) they do not rate the condition **STRONG-4** higher than the condition **FALSE-0**.

3.3 Sample size

We will test 200 participants after exclusions (we will add participants so as to reach this target after excluding participants in a first batch). In a pilot study, we had tested 120 participants on only 3 conditions ((a) one where every box among 10 boxes had exactly one object in it, (b) one where every box among 10 boxes had several objects in it, and (c) one where 5 among 10 boxes had several objects and the other 5 exactly 1), and we tested the pairwise differences ((a) vs (c) and (b) vs (c) (testing with a t.test comparing the marginal means given a mixed effect linear model that fits all the data, using the R package emmeans). We estimated that we needed about 100 participants to obtain 80% power for the weaker effect if we replicated exactly this experiment. But we now introduce intermediate conditions and will perform more fine-grained tests to assess gradience (as explained below), so we computed power based on half of the effect size we observed. As each participant in the present study will see many more trials, we estimate that a sample size of 200 participants will reach the required statistical power.

4. Hypotheses

Within false cases, we expect a gradient effect, such that average cursor judgments will correlate positively with the proportion of strong verifier boxes and negatively with the proportion of empty boxes.

Within true cases, we also expect a gradient effect, such that average cursor judgments will correlate positively with the proportion of strong verifier boxes and negatively with the proportion of boxes containing a unique circle.

5. Statistical analyses

1. We will subset the literally true conditions (LITERAL, WEAK and STRONG) and run a mixed effect linear model predicting the responses as a function of the number of strong verifiers (a factor coded as numeric), with a random intercept and random slope per participant. In case of failure of convergence or singular fit, we will remove the random slope. We will compare this model with a null model which will just have the intercept and with the same random effects structure by means of a LRT test. We will that there is evidence for an effect of the number of strong verifiers if the p-value for the LRT test is <0.05
2. We will also run the very same model (with maximal random effects structure) on the WEAK conditions, to determine whether within the weak conditions, the number of strong verifiers plays a role, and likewise compare it to the null model by means of an LRT test.
3. We will correct for multiple comparisons using the Holm-Bonferroni method.
4. We will also run a number of exploratory analyses in order to have a fine-grained understanding of whatever gradient effects are found and how they might be linked to the truth-value of different readings in different situations.

Plural expressions and cumulativity: a comprehension study.

Benjamin Spector, Claire Rong

1. Background

There exist several approaches to the semantic and pragmatic interpretation of plural indefinites: the implicature approach (Sauerland 2003, Spector 2007, Zweig 2007), which comes in many different variety, the homogeneity approach (Kriz 2017), the presuppositional approach (a version of which is also an implicature-based approach, Bassi, del Pinal & Sauerland 2021). These accounts all agree that the multiplicity inference triggered by plural indefinites is not a standard entailment. They all predict that the multiplicity inference is not part of the content that is negated in a simple negative sentence. Theories differ from each other (even when they belong to the same class of approaches) with respect to cases where a plural indefinite is under the scope of a universal quantifier.

We have run some exploratory experiments and in production we observed gradient effects regarding the choice between sentences like “Every box contains pencils” vs “Every box contains a pencil”, depending on the proportion of boxes that contain one or several pencils (cf. preregistration <https://osf.io/5dxe7>).

We then investigated gradience in comprehension (cf. preregistration <https://osf.io/7dkfn>) - a gradient effect whereby the proportion of boxes with several objects influence behavior in a truth-value judgment task is not predicted by existing theories (except maybe Enguehard 2024). We observed that in a uniformly singular situation (i.e. each box containing exactly one object), the average judgment was over 70% in favour of “Each box contains [bare plural]” being a “good description” of the situation. We now aim to check whether this judgment could have resulted from a marginal cumulative interpretation (supposedly less accessible with “each” than with “every”, but maybe still present).

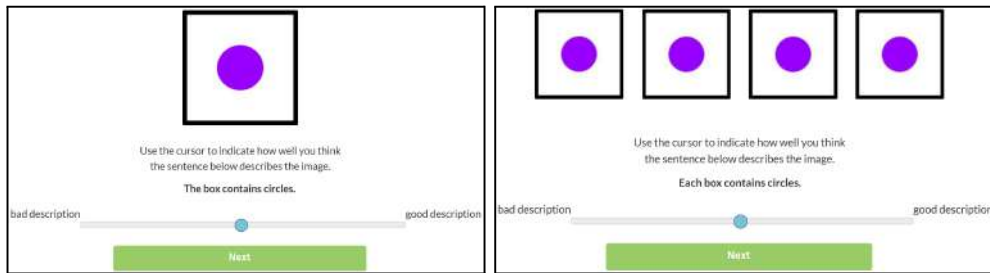
2. Study design

2.1 Participants

We will test adult participants on the Prolific platform who report English to be their ‘first language’, ‘primary language’ and ‘earliest language in life’.

2.2 Stimuli and procedure

A trial consists of a picture and a sentence: the sentence follows a fixed structure “[Each/the] box contains [plural expression]” and is paired with a picture made up of either 1 or 4 boxes containing exactly one circle. Here are two examples of trials:



There are two factors and a total of six conditions in the experiment:

F1. The plural expression is one of the following: “several NPs”, “some NPs”, or a bare plural.

F2. The type of picture shown: (with each box always containing a unique circle.)

- either with 1 box, in which case the sentence starts with “The box”
- or with 4 boxes, in which case the sentence starts with “Each box”.

Each trial is accompanied by the same instruction: “Use the cursor to indicate how well you think the sentence below describes the image”. The cursor moves on a continuous scale ranging from “bad description” (left extremity) to “good description” (right extremity). Participants’ cursor ratings are saved as integers between 1 and 100, but the rating number is not visible to participants themselves. Each participant completes exactly one trial, so there are six different groups. Assignment to a group is randomized.

3. Sampling plan

3.1 Data collection procedure

Participants will be recruited through the online platform Prolific and paid £0.15 for their participation.

3.2 Inclusion and exclusion criteria

Participants will be excluded from the analyses if they do not complete the full experiment.

3.3 Sample size

We will test 600 participants in total, with approximately 100 participants per condition, given random assignment.

In a pilot study, we tested between 20 and 25 participants per condition. We examined the three p-values corresponding to three interaction contrasts, i.e. comparisons of the effect of box number (1 vs. 4) between each pair of sentence types. Based on the pilot data, we conducted a power analysis which indicated that a sample size of 80 participants per condition would yield a statistical power of approximately 80% to detect all three interaction effects. To allow for participant exclusions and ensure conservative coverage, we will slightly exceed this threshold and recruit 100 participants per condition.

4. Hypotheses

If the high ratings observed in the literal condition of our previous experiment (<https://osf.io/7dkfn>) are due to the marginal availability of a cumulative interpretation when multiple boxes are present, we expect picture type (F2) to significantly influence participants' responses. Specifically, we predict that sentences will receive higher scores when paired with 4-box images compared to 1-box images, across all three sentence types (F1).

While we do expect “some NPs” and bare plurals to be more acceptable across the board when there is only one referent per box (for both 1-box and 4-box cases), we do not expect, on theoretical grounds, that the availability of a cumulative interpretation (which would allow people to like the 4-box condition even when they derive a multiplicity inference) would itself depend on the form of the indefinite. In other terms, we do not expect, on theoretical grounds, to observe significant interactions between the two factors. However, it is also possible that ‘inclusive’ plural items (like “some NPs” and bare plurals) are more likely to license such a cumulative interpretation with *each*, in which case we will observe interaction effects.

5. Statistical analyses

We will fit a linear model with F1, F2, and their interaction ($F1 \times F2$) as predictors (i.e., $\text{score} \sim 1 + F1 + F2 + F1:F2$). For each level of F1, we will compare the 4-box and 1-box conditions by applying a Wald t-test to the estimated marginal means obtained from the model, using the emmeans package.

In addition, we will test for interaction effects between the 1-box/4-box manipulation and each pairwise comparison among the three levels of F1 (bare plural, some, several), resulting in three distinct 2×2 interactions. These interaction effects will be tested by computing differences of differences (i.e., 2×2 interaction contrasts) based on the model-predicted marginal means. Each contrast will be evaluated using a Wald t-test, implemented via the emmeans package.

For balanced designs like ours, these tests performed using emmeans are mathematically equivalent to testing specific linear combinations of model coefficients and yield the same p-values as those obtained from the corresponding interaction terms in a linear model, provided the factors are appropriately coded to reflect the contrasts of interest.

Since we are running six Wald t-tests, we will correct for multiple comparisons using the Holm-Bonferroni method.

“several NPs” in universally quantified statements: a comprehension study.

Benjamin Spector, Claire Rong

1. Background

There exist several approaches to the semantic and pragmatic interpretation of plural indefinites: the implicature approach (Sauerland 2003, Spector 2007, Zweig 2007), which comes in many different varieties, the homogeneity approach (Kriz 2017), the presuppositional approach (a version of which is also an implicature-based approach, Bassi, del Pinal & Sauerland 2021). These accounts all agree that the multiplicity inference triggered by plural indefinites is not a standard entailment. They all predict that the multiplicity inference is not part of the content that is negated in a simple negative sentence. Theories differ from each other (even when they belong to the same class of approaches) with respect to cases where a plural indefinite is under the scope of a universal quantifier.

We have run some exploratory production experiments and we observed gradient effects regarding the choice between sentences like *Every box contains pencils* vs *Every box contains a pencil*, depending on the proportion of boxes that contain one or several pencils (cf. preregistration <https://osf.io/5dxe7>).

We then investigated gradience in comprehension of “Each box contains [bare plural]” (cf. preregistration <https://osf.io/7dkfn>) - a gradient effect whereby the proportion of boxes with several objects influences behavior in a truth-value judgment task, which is not predicted by existing theories (except maybe Enguehard 2024). We want to see whether similar gradient effects are found with “several” (“Each box contains [several NPs]”), as it is a strictly plural expression and no pragmatic strengthening takes place.

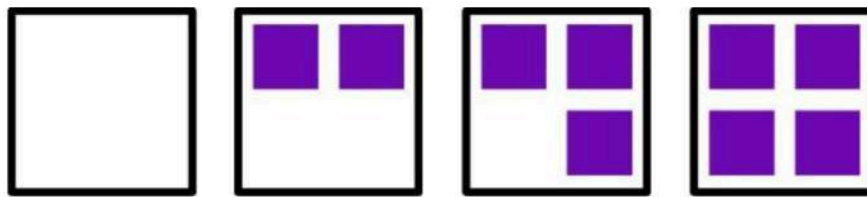
2. Study design

2.1 Participants

We will test adult participants who report English to be their ‘first language’, ‘primary language’ and ‘earliest language in life’.

2.2 Stimuli and procedure

A trial consists of a picture and a sentence: the sentence follows a fixed structure “Each box contains [several NPs]” and is paired with a picture made up of 4 boxes containing between 0 and 4 geometric shapes. Here is an example of a trial:



Use the cursor to indicate how well you think
the sentence below describes the image.

Each box contains several squares.



There are two types of conditions in the experiment:

1. Literal truth conditions

- A sentence is literally true (in what follows, we will also call it *strong*) if all boxes contain several shapes (1 condition).
- A sentence is literally false otherwise (8 conditions).

2. Experimental conditions

- A box containing multiple shapes is called a strong verifier.
- Conditions are labeled based on the reading they satisfy (FALSE, STRONG), followed by the number of empty boxes, then by the number of strong verifiers in the image.

stimuli pictures (examples with circles)	label for condition
	STRONG-0-4
	FALSE-0-3
	FALSE-0-2
	FALSE-0-1
	FALSE-0-0

	FALSE-1-3
	FALSE-2-2
	FALSE-3-1
	FALSE-4-0

Each trial is accompanied by the same instruction: “Use the cursor to indicate how well you think the sentence below describes the image”. In the sentence “Each box contains [several NPs]”, NPs was one of the three following words, always corresponding to the geometric shape shown in the picture: *circles*, *triangles*, or *squares*. The cursor moves on a continuous scale ranging from “bad description” (left extremity) to “good description” (right extremity). Participants’ cursor ratings are saved as integers between 1 and 100, but the rating number is not visible to participants themselves.

Each participant completes 27 randomized trials, with each condition appearing three times, using three different geometric shapes in three different colors (coupling of shapes/colors is reported in the table below). The stimuli are available in the stimuli folder (in this OSF project).

	green circles	orange circles	purple circles	green squares	orange squares	purple squares	green triangles	orange triangles	purple triangles
FALSE-1		x		x					x
FALSE-2			x		x		x		
FALSE-3	x					x		x	
LITERAL-0		x		x					x
WEAK-1			x		x		x		
WEAK-2	x					x		x	
WEAK-3		x		x					x
STRONG-4			x		x		x		

The final page contains an attention check.

3. Sampling plan

3.1 Data collection procedure

Participants will be recruited through the online platform Prolific and paid £0.60 for their participation.

3.2 Inclusion and exclusion criteria

Participants will be excluded from the analyses if:

- i) they do not complete the full experiment;
- ii) they fail the attention check;
- iii) they do not rate the condition **STRONG-0-4** higher than the condition **FALSE-4-0**.

3.3 Sample size

We will test 70 participants after exclusions, based on power estimates using the data collected through the previous experiment with bare plurals (this corresponds to power $[1-\beta] > 90\%$)

4. Hypotheses

Within at-least-one-empty-box cases, we expect a gradient effect, such that average cursor judgments will correlate positively with the proportion of strong verifier boxes and negatively with the proportion of empty boxes, even if the picture makes the sentence false.

Within no-empty-box cases, we also expect a gradient effect, such that average cursor judgments will correlate positively with the proportion of strong verifier boxes and negatively with the proportion of boxes containing a unique circle.

5. Statistical analyses

1. We first fit a mixed-effects model to the subset of no-empty-box conditions, excluding STRONG-0-4, predicting cursor score as a function of the number of strong verifiers (i.e. number of boxes with more than one item), with a random intercept and random slope per participant. In case of failure of convergence or singular fit, we will remove the random slope. We will compare this model with a null model which will just have the intercept and with the same random effects structure by means of a LRT test. We will take a p-value < 0.05 from the LRT test as evidence for an effect of the number of strong verifiers.
2. We do the same for the subset of at-least-one-empty-box conditions.
3. We will correct for multiple comparisons using the Holm-Bonferroni method.

“Some NPs” in universally quantified statements: a comprehension study.

Benjamin Spector, Claire Rong

1. Background

There exist several approaches to the semantic and pragmatic interpretation of plural indefinites: the implicature approach (Sauerland 2003, Spector 2007, Zweig 2007), which comes in many different variety, the homogeneity approach (Kriz 2017), the presuppositional approach (a version of which is also an implicature-based approach, Bassi, del Pinal & Sauerland 2021). These accounts all agree that the multiplicity inference triggered by plural indefinites is not a standard entailment. They all predict that the multiplicity inference is not part of the content that is negated in a simple negative sentence. Theories differ from each other (even when they belong to the same class of approaches) with respect to cases where a plural indefinite is under the scope of a universal quantifier.

We have run some exploratory experiments and in production we observed gradient effects regarding the choice between sentences like *Every box contains pencils* vs *Every box contains a pencil*, depending on the proportion of boxes that contain one or several pencils (cf. preregistration <https://osf.io/5dxe7>).

We then investigated gradience in comprehension (cf. preregistration <https://osf.io/7dkfn>) - a gradient effect whereby the proportion of boxes with several objects influence behavior in a truth-value judgment task, which is not predicted by existing theories (except maybe Enguehard 2024). We observed that in a uniformly singular situation (i.e. each box containing exactly one object), the average judgment was over 70% in favour of “Each box contains [bare plural]” being a “good description” of the situation. We also conducted a pilot comprehension study using the same uniformly singular situation, but this time with the sentence “Each box contains [some NPs]”. We found that average judgment scores were lower with “some NPs” than with bare plurals. Based on this, we plan to replicate the design of the previous comprehension study (<https://osf.io/7dkfn>), replacing bare plurals with “some NPs”. Our goal is to observe greater gradience within the literally true conditions: because “some NPs” received lower scores in uniformly singular contexts, there is more room for scores to increase as the number of strong plural verifiers grows.

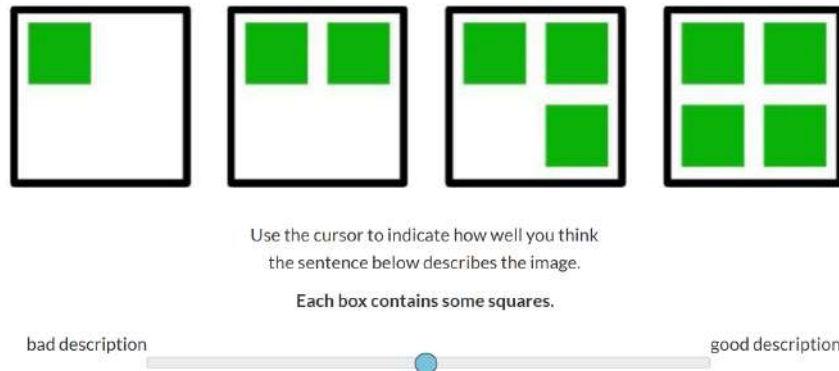
2. Study design

2.1 Participants

We will test adult participants who report English to be their ‘first language’, ‘primary language’ and ‘earliest language in life’.

2.2 Stimuli and procedure

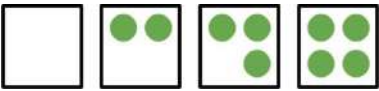
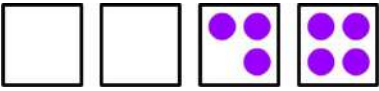
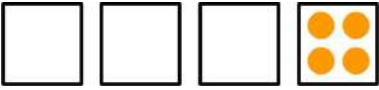
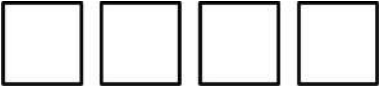
A trial consists of a picture and a sentence: the sentence follows a fixed structure “Each box contains [some NPs]” and is paired with a picture made up of 4 boxes containing between 0 and 4 geometric shapes. Here is an example of a trial:



There are two types of conditions in the experiment:

1. Literal truth conditions
 - A sentence is literally false if at least one box is empty (4 conditions).
 - A sentence is literally true otherwise (5 conditions).
2. Experimental conditions
 - A box containing multiple shapes is called a strong verifier.
 - Conditions are labeled based on the strongest reading they satisfy (FALSE, LITERAL, WEAK, or STRONG), followed by the number of strong verifiers in the image.

stimuli pictures (examples with circles)	label for condition	readings
	STRONG-4	true for STRONG+WEAK+LITERAL readings
	WEAK-3	true for WEAK+LITERAL readings
	WEAK-2	true for WEAK+LITERAL readings
	WEAK-1	true for WEAK+LITERAL readings
	LITERAL-0	true for LITERAL reading

	FALSE-3	false for LITERAL reading
	FALSE-2	false for LITERAL reading
	FALSE-1	false for LITERAL reading
	FALSE-0	false for LITERAL reading

Each trial is accompanied by the same instruction: “Use the cursor to indicate how well you think the sentence below describes the image”. In the sentence “Each box contains [some NPs]”, NPs was one of the three following words, always corresponding to the geometric shape shown in the picture: *circles*, *triangles*, or *squares*. The cursor moves on a continuous scale ranging from “bad description” (left extremity) to “good description” (right extremity). Participants’ cursor ratings are saved as integers between 1 and 100, but the rating number is not visible to participants themselves.

Each participant completes 27 randomized trials, with each condition appearing three times, using three different geometric shapes in three different colors (coupling of shapes/colors is reported in the table below). The stimuli are available in the stimuli folder (in this OSF project).

	green circles	orange circles	purple circles	green squares	orange squares	purple squares	green triangles	orange triangles	purple triangles
FALSE-1		x		x					x
FALSE-2			x		x		x		
FALSE-3	x					x		x	
LITERAL-0		x		x					x
WEAK-1			x		x		x		
WEAK-2	x					x		x	
WEAK-3		x		x					x
STRONG-4			x		x		x		

The final page contains an attention check.

3. Sampling plan

3.1 Data collection procedure

Participants will be recruited through the online platform Prolific and paid £0.60 for their participation.

3.2 Inclusion and exclusion criteria

Participants will be excluded from the analyses if:

- i) they do not complete the full experiment;
- ii) they fail the attention check;
- iii) they do not rate the condition **STRONG-4** higher than the condition **FALSE-0**.

3.3 Sample size

We will test 200 participants after exclusions, based on power estimates using the data collected through the previous experiment with bare plurals.

4. Hypotheses

Within false cases, we expect a gradient effect, such that average cursor judgments will correlate positively with the proportion of strong verifier boxes and negatively with the proportion of empty boxes.

Within true cases, we also expect a gradient effect, such that average cursor judgments will correlate positively with the proportion of strong verifier boxes and negatively with the proportion of boxes containing a unique circle.

5. Statistical analyses

1. We will analyze the data using linear mixed-effects models to examine how cursor-based truth-value judgments are influenced by multiple factors. We define four predictors:
 - c_vrf (the number of strong verifiers in a picture, a factor coded as numeric)
 - c_lit (a binary variable indicating whether a condition is literally true)
 - c_weak (indicating whether the condition supports a weak reading)
 - c_str (indicating whether the condition supports a strong reading).
2. We first fit a mixed-effects model to the subset of literally true conditions (LITERAL, WEAK and STRONG) predicting cursor score as a function of c_vrf, with a random intercept and random slope per participant. In case of failure of convergence or singular fit, we will remove the random slope. We will compare this model with a null model which will just have the intercept and with the same random effects structure by means of a

LRT test. We will take a p-value <0.05 from the LRT test as evidence for an effect of the number of strong verifiers.

3. We will also run the very same model (with maximal random effects structure) on the WEAK conditions, to determine whether within the weak conditions, the number of strong verifiers plays a role, and likewise compare it to the null model by means of an LRT test.
4. We will correct for multiple comparisons using the Holm-Bonferroni method.
5. Finally, we will perform model comparison across all possible combinations of the four predictors by fitting 16 models, each including a different subset of the predictors, and ranking them using the Bayesian Information Criterion (BIC), using as a random effect structure one with only a random intercept per participant, so as to ensure convergence of as many models as possible (we may explore also models with random slopes). The best-fitting model will be selected based on the lowest BIC. We will consider a model to be clearly better than the next-best model if the BIC difference (ΔBIC) exceeds 2, following standard guidelines (e.g., Raftery, 1995).

“Some NPs” in universally quantified statements: a comprehension study. Binary response task version.

Benjamin Spector, Claire Rong

1. Background

There exist several approaches to the semantic and pragmatic interpretation of plural indefinites: the implicature approach (Sauerland 2003, Spector 2007, Zweig 2007), which comes in many different variety, the homogeneity approach (Kriz 2017), the presuppositional approach (a version of which is also an implicature-based approach, Bassi, del Pinal & Sauerland 2021). These accounts all agree that the multiplicity inference triggered by plural indefinites is not a standard entailment. They all predict that the multiplicity inference is not part of the content that is negated in a simple negative sentence. Theories differ from each other (even when they belong to the same class of approaches) with respect to cases where a plural indefinite is under the scope of a universal quantifier.

We have run some exploratory experiments and we observed gradient effects in production regarding the choice between sentences like *Every box contains pencils* vs *Every box contains a pencil*, depending on the proportion of boxes that contain one or several pencils (cf. preregistration <https://osf.io/5dxe7>).

We then investigated gradience in comprehension (cf. preregistrations <https://osf.io/7dkfn> and <https://osf.io/ys9wb>), i.e. a gradient effect whereby the proportion of boxes with several objects influences behavior in a truth-value judgment task, which is not predicted by existing theories (except maybe Enguehard 2024). Our two previous comprehension studies respectively asked for judgments of “Each box contains [bare plural]” and “Each box contains [some NPs]”. As an exploratory analysis, we binarized continuous slider responses by applying a threshold at 50 (on a 0-100 scale), treating values above this point as “true” judgments. The resulting simulated binary data still displayed gradient patterns.

In the current study, we aim to determine whether such gradient effects are still present when participants are given an explicit binary forced-choice task. We focus on “some NPs” sentences rather than bare plurals because, in our previous data, “some NPs” received lower average scores in uniformly singular scenarios (i.e., each box containing exactly one object). This suggests that “some NPs” provide greater room to observe gradient effects within scenarios that are literally true.

2. Study design

2.1 Participants

We will test adult participants located in the US or in the UK who report English to be their ‘first language’, ‘primary language’ and ‘earliest language in life’.

2.2 Stimuli and procedure

A trial consists of a picture and a sentence: the sentence follows a fixed structure “Each box contains [some NPs]” and is paired with a picture made up of 4 boxes containing between 0 and 4 geometric shapes. Here is an example of a trial:



Do you think the sentence below is true or false?

Each box contains some squares.

☐ false

☐ true

There are two types of conditions in the experiment:

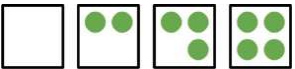
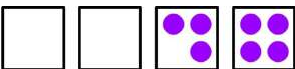


1. Literal truth conditions

- A sentence is literally false if at least one box is empty (4 conditions).
- A sentence is literally true otherwise (5 conditions).

2. Experimental conditions

- A box containing multiple shapes is called a strong verifier.
- Conditions are labeled based on the strongest reading they satisfy (FALSE, LITERAL, WEAK, or STRONG), followed by the number of strong verifiers in the image.

stimuli pictures (examples with circles)	label for condition	readings
	STRONG-4	true for STRONG+WEAK+LITERAL readings
	WEAK-3	true for WEAK+LITERAL readings
	WEAK-2	true for WEAK+LITERAL readings
	WEAK-1	true for WEAK+LITERAL readings
	LITERAL-0	true for LITERAL reading

	FALSE-3	false for LITERAL reading
	FALSE-2	false for LITERAL reading
	FALSE-1	false for LITERAL reading
	FALSE-0	false for LITERAL reading

Each trial is accompanied by the same question: “Do you think the sentence below is true or false?”. In the sentence “Each box contains [some NPs]”, NPs was one of the three following words, always corresponding to the geometric shape shown in the picture: *circles*, *triangles*, or *squares*. The binary choice task offers the options “false” on the left and “true” on the right.

Each participant completes 27 randomized trials, with each condition appearing three times, using three different geometric shapes in three different colors (coupling of shapes/colors is reported in the table below). The stimuli are available in the stimuli folder (in this OSF project).

	green circles	orange circles	purple circles	green square s	orange square s	purple square s	green triangle s	orange triangle s	purple triangle s
FALSE-1		x		x					x
FALSE-2			x		x		x		
FALSE-3	x					x		x	
LITERAL-0		x		x					x
WEAK-1			x		x		x		
WEAK-2	x					x		x	
WEAK-3		x		x					x
STRONG-4			x		x		x		

The final page contains an attention check.

3. Sampling plan

3.1 Data collection procedure

Participants will be recruited through the online platform Prolific and paid £0.60 for their participation.

3.2 Inclusion and exclusion criteria

Participants will be excluded from the analyses if:

- i) they do not complete the full experiment;
- ii) they fail the attention check;
- iii) they do not rate every trial of the condition **FALSE-0** as false.

3.3 Sample size

We will test 200 participants after exclusions, using the same sample size as the previous version of the study that used a continuous cursor response option (<https://osf.io/ys9wb>).

4. Hypotheses

Within false cases, we expect a gradient effect, such that the proportion of “true” responses will correlate positively with the proportion of strong verifier boxes and negatively with the proportion of empty boxes.

Within true cases, we also expect a gradient effect, such that the proportion of “true” responses will correlate positively with the proportion of strong verifier boxes and negatively with the proportion of boxes containing a unique circle.

5. Statistical analyses

1. We will analyze the data using logistic mixed-effects models to examine how binary truth-value judgments are influenced by various factors. We define four predictors:
 - c_vrf: the number of strong verifiers in a picture (numeric)
 - c_lit: whether a condition is literally true (binary)
 - c_weak: whether the condition supports a weak reading (binary)
 - c_str: whether the condition supports a strong reading (binary).
2. We will first fit a logistic mixed-effects model to the subset of literally true conditions (LITERAL, WEAK, and STRONG), predicting the probability of a “true” response as a function of c_vrf, with a random intercept and random slope for c_vrf by participant. In case of convergence issues or singular fit, the random slope will be removed. We will compare this model to a null model (intercept only, with the same random-effects

structure) using a likelihood ratio test (LRT). A p-value <0.05 will be taken as evidence for a significant effect of the number of strong verifiers.

3. We will also run the very same model (with maximal random effects structure) on the WEAK conditions, to determine whether within the weak conditions, the number of strong verifiers plays a role, and likewise compare it to the null model by means of an LRT test.
4. We will correct for multiple comparisons using the Holm-Bonferroni method.
5. Finally, we will perform model comparison across all possible combinations of the four predictors by fitting 16 models, each including a different subset of the predictors, and ranking them using the Bayesian Information Criterion (BIC), using as a random effect structure one with only a random intercept per participant, so as to ensure convergence of as many models as possible (we may explore also models with random slopes). The best-fitting model will be selected based on the lowest BIC. We will consider a model to be clearly better than the next-best model if the BIC difference (ΔBIC) exceeds 2, following standard guidelines (e.g., Raftery, 1995).

“yixie NP” in universally quantified statements: a comprehension study.

Benjamin Spector, Claire Rong

This is an updated version of the preregistration originally posted at <https://osf.io/nv8ms>. Due to a technical issue on the OSF platform, we were unable to use the standard “update” procedure to modify the original preregistration. The updated section is 3.1 Data Collection Procedure.

1. Background

There exist several approaches to the semantic and pragmatic interpretation of English plural indefinites: the implicature approach (Sauerland 2003, Spector 2007, Zweig 2007), which comes in many different variety, the homogeneity approach (Kriz 2017), the presuppositional approach (a version of which is also an implicature-based approach, Bassi, del Pinal & Sauerland 2021). These accounts all agree that the multiplicity inference triggered by plural indefinites is not a standard entailment. They all predict that the multiplicity inference is not part of the content that is negated in a simple negative sentence. Theories differ from each other (even when they belong to the same class of approaches) with respect to cases where a plural indefinite is under the scope of a universal quantifier. In Chinese, the quantifier —些 (yīxiē) used with plural indefinites also triggers a multiplicity inference without it being a standard entailment.

We have run some exploratory experiments and in production we observed gradient effects regarding the choice between sentences like *Every box contains pencils* vs *Every box contains a pencil*, depending on the proportion of boxes that contain one or several pencils (cf. preregistration <https://osf.io/5dxe7>).

We then investigated gradience in comprehension of plural indefinites in English (cf. preregistrations <https://osf.io/7dkfn> and <https://osf.io/ys9wb>). We observed a gradient effect whereby the proportion of boxes with several objects influences behavior in a truth-value judgment task, which is not predicted by existing theories (except maybe Enguehard 2024). Our two previous English comprehension studies respectively asked for judgments of “Each box contains [bare plural]” and “Each box contains [some NPs]”. The current study aims to examine whether similar gradient effects are observed in Chinese, using the quantifier yīxiē in place of “some”.

2. Study design

2.1 Participants

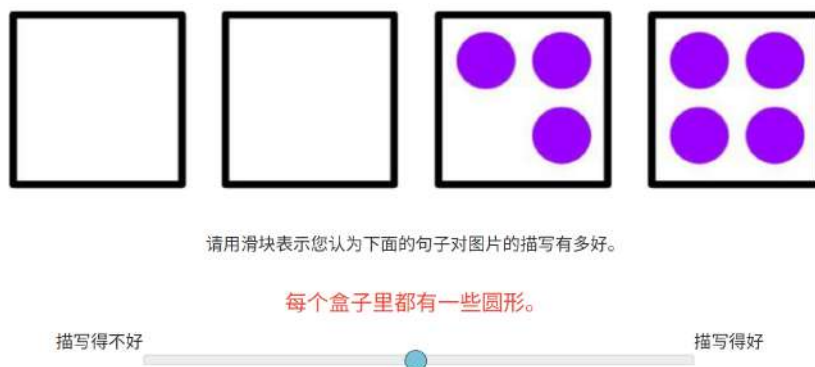
We will test adult participants who report Chinese to be their ‘primary language’ and ‘earliest language in life’.

2.2 Stimuli and procedure

A trial consists of a picture and a sentence: the sentence follows a fixed structure 每个盒子里都有一些[NP] (equivalent in the English experiment: “Each box contains [some NPs]”) and is paired with a picture made up of 4 boxes containing between 0 and 4 geometric shapes. Here is a glossed example:

měi gè hé-zi lǐ dōu yǒu yī xiē
 每 个 盒 子 里 都 有 一 些 [NP]
 each CL box in DIST have one xie [NP]

Here is an example of a trial:



There are two types of conditions in the experiment:

1. Literal truth conditions
 - A sentence is literally false if at least one box is empty (4 conditions).
 - A sentence is literally true otherwise (5 conditions).
2. Experimental conditions
 - A box containing multiple shapes is called a strong verifier.
 - Conditions are labeled based on the strongest reading they satisfy (FALSE, LITERAL, WEAK, or STRONG), followed by the number of strong verifiers in the image.

stimuli pictures (examples with circles)	label for condition	readings
	STRONG-4	true for STRONG+WEAK+LITERAL readings
	WEAK-3	true for WEAK+LITERAL readings

	WEAK-2	true for WEAK+LITERAL readings
	WEAK-1	true for WEAK+LITERAL readings
	LITERAL-0	true for LITERAL reading
	FALSE-3	false for LITERAL reading
	FALSE-2	false for LITERAL reading
	FALSE-1	false for LITERAL reading
	FALSE-0	false for LITERAL reading

Each trial is accompanied by this instruction translated into Chinese: “Use the cursor to indicate how well you think the sentence below describes the image”. In the sentence “Each box contains [yixie NP]”, NPs was one of the three following words, always corresponding to the geometric shape shown in the picture: *circles*, *triangles*, or *squares*. The cursor moves on a continuous scale ranging from “描写得不好” (“describes badly”, left extremity) to “描写得好” (“describes well”, right extremity). Participants’ cursor ratings are saved as integers between 1 and 100, but the rating number is not visible to participants themselves.

Each participant completes 27 randomized trials, with each condition appearing three times, using three different geometric shapes in three different colors (coupling of shapes/colors is reported in the table below). The stimuli are available in the stimuli folder (in this OSF project).

	green circles	orange circles	purple circles	green squares	orange squares	purple squares	green triangles	orange triangles	purple triangles
FALSE-1		x		x					x
FALSE-2			x		x		x		
FALSE-3	x					x		x	
LITERAL-0		x		x					x

WEAK-1			x		x		x		
WEAK-2	x					x		x	
WEAK-3		x		x					x
STRONG-4			x		x		x		

Before starting the experimental trials, participants will answer the question (translated into Chinese) “Is Chinese your native language?”, with a binary response choice.

The final page contains an attention check.

3. Sampling plan

3.1 Data collection procedure

Part of our participants had first been recruited through direct contact and snowball sampling (based on voluntary participation). This method yielded only 23 sets of responses (before exclusions). This motivated an update of this preregistration with a change in recruitment methods.

Remaining participants will be recruited through the online platform Prolific and paid £0.60 for their participation.

3.2 Inclusion and exclusion criteria

Participants will be excluded from the analyses if:

- i) they do not complete the full experiment;
- ii) they answer “no” at the question “Is Chinese your native language?”
- iii) they fail the attention check;
- iv) they do not rate every trial of the condition **STRONG-4** higher than every trial of the condition **FALSE-0**.

3.3 Sample size

We will test 200 participants after exclusions, based on power estimates using the data collected through the previous experiment with English bare plurals.

4. Hypotheses

Within false cases, we expect a gradient effect, such that average cursor judgments will correlate positively with the proportion of strong verifier boxes and negatively with the proportion of empty boxes.

Within true cases, we also expect a gradient effect, such that average cursor judgments will correlate positively with the proportion of strong verifier boxes and negatively with the proportion of boxes containing a unique circle.

5. Statistical analyses

1. We will analyze the data using linear mixed-effects models to examine how cursor-based truth-value judgments are influenced by multiple factors. We define four predictors:
 - c_vrf (the number of strong verifiers in a picture, a factor coded as numeric)
 - c_lit (a binary variable indicating whether a condition is literally true)
 - c_weak (indicating whether the condition supports a weak reading)
 - c_str (indicating whether the condition supports a strong reading).
2. We first fit a mixed-effects model to the subset of literally true conditions (LITERAL, WEAK and STRONG) predicting cursor score as a function of c_vrf, with a random intercept and random slope per participant. In case of failure of convergence or singular fit, we will remove the random slope. We will compare this model with a null model which will just have the intercept and with the same random effects structure by means of a LRT test. We will take a p-value <0.05 from the LRT test as evidence for an effect of the number of strong verifiers.
3. We will also run the very same model (with maximal random effects structure) on the WEAK conditions, to determine whether within the weak conditions, the number of strong verifiers plays a role, and likewise compare it to the null model by means of an LRT test.
4. We will correct for multiple comparisons using the Holm-Bonferroni method.
5. Finally, we will perform model comparison across all possible combinations of the four predictors by fitting 16 models, each including a different subset of the predictors, and ranking them using the Bayesian Information Criterion (BIC), using as a random effect structure one with only a random intercept per participant, so as to ensure convergence of as many models as possible (we may explore also models with random slopes). The best-fitting model will be selected based on the lowest BIC. We will consider a model to be clearly better than the next-best model if the BIC difference (ΔBIC) exceeds 2, following standard guidelines (e.g., Raftery, 1995).