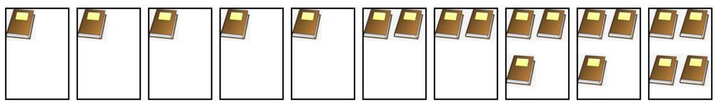
Plurals under universal quantification: disentangling readings from gradient effects

Claire RONG & Benjamin SPECTOR
Institut Jean Nicod (CNRS, ENS-PSL, EHESS), École Normale Supérieure, Paris

Main problem

(1) Each box contains books.



Consider a situation where some but not all boxes contain ≥ 2 books, others contain $= 1$ book (example pictured).
Theories differ on the truth-values of (1) in this situation.
An intermediate truth-value judgment does not necessarily reflect a distinct reading.

> **Gradient effects:** ratings vary continuously with the similarity between a scenario and one that makes the sentence true, even when the same reading is satisfied.

→ can lead to **spurious inferences**: e.g. Stateva et al. 2016 interpreted intermediate ratings in such scenarios as evidence for an intermediate reading, but did not control for gradient effects.
→ **How can we disentangle readings from potential gradient effects?**

Plurals under universal quantification: theoretical predictions

Plural morphology does not always convey strict plurality (≥ 2).
While plural indefinites and bare plurals typically trigger a **multiplicity inference** in upward-entailing (UE) environments, strict plurality meaning is usually not present in downward-entailing (DE) environments:

(2) The box contains books. → box contains ≥ 2 books.

(3) The box doesn't contain books. → ¬ (box contains ≥ 1 book) ≡ box contains 0 book.

Puzzle of a **logical gap**: the meaning of bare plurals in UE environments is not the negation of their meaning in DE environments.
(2) has two potential readings: 'the box contains ≥ 2 books', 'the box contains ≥ 1 book'.
There are potentially even **more readings in a universally quantified sentence**:

(1) Each box contains books.

Three theoretically possible readings of (1):

1. **Literal reading:** each box contains ≥ 1 book.
2. **Weak reading:** each box contains ≥ 1 book, and it is not the case that all contain $= 1$ book.
3. **Strong reading:** each box contains ≥ 2 books.

Logical strength: strong > weak > literal.
Theories generate different sets of readings, and therefore make different predictions on the truth-value of (1) in **mixed scenarios**, where some but not all boxes contain ≥ 2 books, others contain $= 1$ book.

Sets of readings predicted:
Spector 2007 → literal, weak, strong
Zweig 2009 → literal, strong
Križ 2017; Bassi et al. 2021 → literal, strong

THE GRADIENT EFFECT CONFOUND:

Even if only **literal + strong** readings exist, mixed scenarios may receive intermediate ratings because they are **closer to a situation satisfying the strong reading**, compared to scenarios where each box contains exactly one book. Thus, an intermediate rating for a mixed scenario could reflect either:

1. a genuine **weak reading**, or
2. an independent **gradient effect**.

→ **How can we disentangle the two possible sources of intermediate judgments?**

Key idea: compare scenarios that satisfy the **same readings** but differ in their similarity to instantiations of the strong reading. A systematic increase in ratings would reveal **gradience independently of readings.**

RESEARCH QUESTIONS:

1. **Empirically, what are the available readings in English?**
2. **How can gradient effects be disentangled from ambiguity between different truth-conditional meanings?**

Experiment 1: Bare plurals

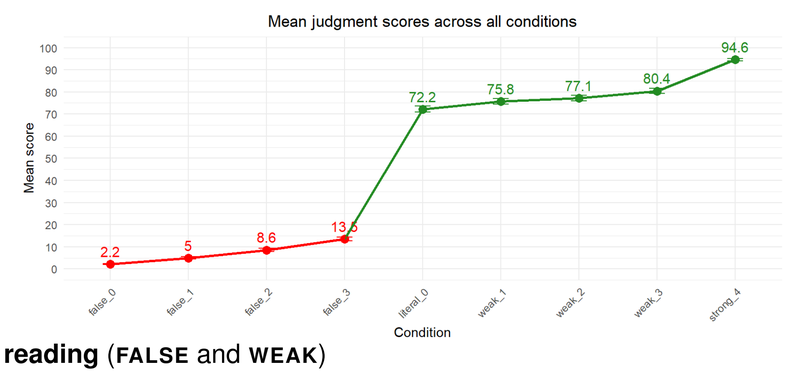
DESIGN

- Picture paired with the sentence "Each box contains [bare plural]" + continuous cursor response (between *good/bad description*).
- **strong verifier** = a box with multiple shapes.
- **Participants:** after exclusions, 200 native English speakers on Prolific.
- **Within-subject design:** each participant saw all 9 conditions 3 times, with different shape/color combinations.
- **Predictors:**
  - c_{vrf} number of strong verifiers
  - c_{lit} literal reading true? (0/1)
  - c_{weak} weak reading true? (0/1)
  - c_{str} strong reading true? (0/1)

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

RESULTS

Linear mixed-effects model: response $\sim c_{vrf} + (1|\text{participant})$ fitted to subset of **WEAK conditions alone**. Compared to a null model through LRT (likelihood ratio test): $\chi^2(1) = 18.75, \quad p < 0.001$.
→ confirms that c_{vrf} significantly improves model fit.



- **Gradience within a same level of reading** (FALSE and WEAK) → not predicted by any theory
- Qualitative shifts from **FALSE to LITERAL**, and from **WEAK to STRONG**, but not from **LITERAL to WEAK** (confirmed by model comparison below)
  → suggests that **only literal and strong readings are accessed**.

MODEL COMPARISON

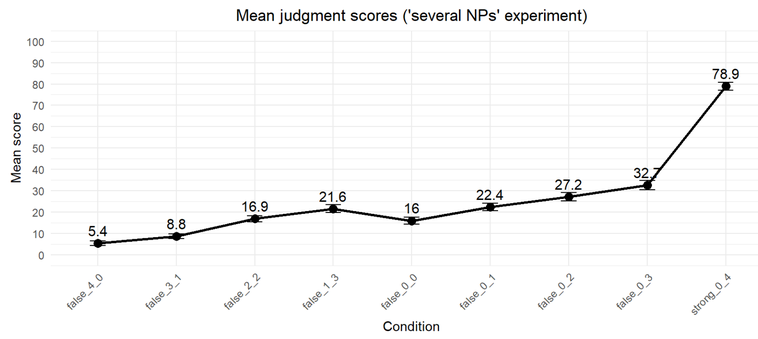
BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion) comparison of $2^4 = 16$ models with all possible combinations of predictors.
Best-fitting model across all 9 conditions **does not include** c_{weak} (indicator that the weak reading is true):

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

$\Delta\text{BIC} = 9$ compared to the second best model, strong evidence for the best model.

Experiment 2: *several NPs*

Replication the design of Exp.1 with "Each box contains [several NPs]".
Only predicted reading: **strong reading** → crucial control for the interpretation of gradient effects.



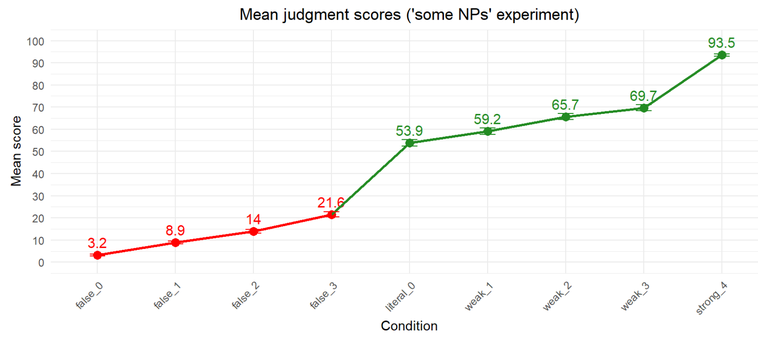
- At-least-one-empty-box conditions: $\chi^2(1) = 380.01, \quad p < 0.001$.
- No-empty-box conditions (excluding the one with 4 verifiers): $\chi^2(1) = 232.31, \quad p < 0.001$.

→ evidence that **gradience is an independent factor**.

DISCUSSION: Gradient effects are observed despite there being **no ambiguity between readings**, suggesting that they **reflect the degree of similarity** to a situation in which the sentence is true.
→ **Gradient effects are not reducible to more readings being satisfied.**
Methodological consequence: detecting semantic readings from graded judgments requires **disentangling reading effects from independent gradient effects.**

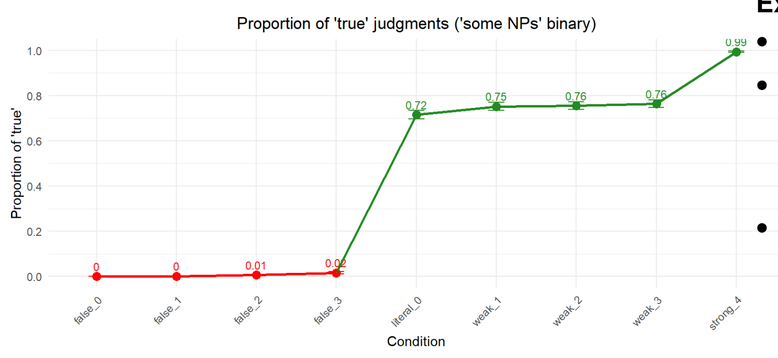
Experiments 3-4: *some NPs*

Same design as Exp. 1-2, with the sentence "Each box contains [some NPs]".
Exp. 3: continuous judgments (as in Exp.1–2). **Exp. 4: binary** truth-value judgments.
200 participants in each experiment, after exclusions.



Exp. 3 (continuous judgments)

- Significant effect of c_{vrf} both in literally true conditions ($\chi^2(1) = 1052.9$, $p < 0.001$) and within **WEAK** conditions ($\chi^2(1) = 65.19$, $p < 0.001$).
- Best model is the same as for bare plurals: it does **not** include c_{weak}.



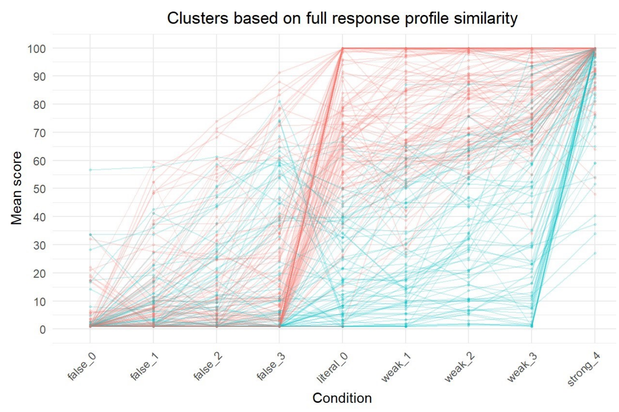
Exp. 4 (binary judgments)

- The same model is ranked first.
- **The weak reading emerges in the best model only in a post-hoc analysis** of the literally-true conditions, and the effect is small.
- Post-hoc analysis using **logistic regression, to be interpreted cautiously:** probabilities close to 0 or 1 can make tiny differences and return large differences in log-odds.

→ **Overall, no qualitative LITERAL–WEAK shift is observed in the judgments.**

Clustering of individual response profiles

METHOD: Rather than averaging across participants, **we represent each participant by a vector of their mean ratings across all conditions** and group participants according to the **similarity of their full response profiles** (hierarchical clustering on Euclidean distances).



DISCUSSION: Across the BP, continuous *some*, and binary *some* datasets, the optimal solution partitions participants into **two clusters**.

- One group shows a larger increase from **FALSE-3 to LITERAL-0**.
- The other shows a larger increase from **WEAK-3 to STRONG-4**.

→ corresponding roughly to participants who favor the **literal** vs. the **strong** reading.
Caveat: clusters are not internally homogeneous and should be interpreted as broad behavioral tendencies rather than categorical speaker types.

Conclusion

1. What are the available readings in English?

- **Literal + strong readings** are supported by model comparisons and clustering, **once gradience is controlled for**.
- Evidence for the **weak reading is inconclusive**: it emerges in a post-hoc analysis of the binary task, but the effect is small and absent from the preregistered analysis.

2. How can gradient effects be disentangled from semantic readings?

- Gradient effects must be controlled for **before inferring semantic readings from intermediate judgments**.
- **Gradient effects are robust:** ratings increase with the number of strong verifiers, even when the same set of readings is satisfied.

References & Acknowledgements

Bassi, Itai, Guillermo Del Pinal, and Uli Sauerland (2021). "Presuppositional exhaustification". Ms.
Križ, Manuel (2017). "Bare plurals, multiplicity, and homogeneity". Ms.
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". In: *Papers dedicated to A.Reboul.*
Zweig, Eytan (2009). "Number-neutral bare plurals and the multiplicity implicature". In: *L & P*.

For useful discussions and feedback, we thank: the anonymous reviewers of SuB and S&P, the audiences of the LINGUAE seminar at ENS and the NihiL seminar at the ILLC. This project has benefited from the support of ANR project ComCog-Mean (ANR-23-CE28-0016; PI: Benjamin Spector).