Two methods to find truth value gaps and their application to the projection problem of homogeneity

Manuel Križ and Emmanuel Chemla

March 18, 2015

Abstract

Presupposition, vagueness, and oddness can lead some sentences to fail to have a clear truth value. The homogeneity property of plural predication with definite descriptions may also create truth-value gaps: *The books are written in Dutch* is true if all relevant books are in Dutch, false if none of them is, and neither true nor false if, say, half of the books are written in Dutch. We study the projection property of homogeneity by deploying methods of general interest to identify truth-value gaps. Method A consists in collecting both truth-value judgments (*completely true* vs. *not completely true*) and, independently, false value judgments (*completely false* vs. *not completely false*). The second method, employed in experiment series B and C, is based on one-shot ternary judgments: *completely true* vs. *completely false* vs. *neither*. After a calibration of these methods, we use them to demonstrate that homogeneity projects out of negation, the scope of universal sentences and the scope of non-monotonic quantifiers such as *exactly two*, to some extent (i.e., in two out of three conceivable kinds of gap situations). We assess our results in light of different theoretical approaches to homogeneity — approaches based on presuppositions, scalar implicatures, and something like supervaluations, respectively. We identify free parameters in these theories and assess various variants of them based on our results. Our experimental paradigms may be of broader significance insofar as they can be applied to other phenomena which result in the failure of a sentence to have a definite truth value.

1 Homogeneity and truth-value gaps

The question of whether a sentence is true or false in a given situation has been part of the foundation of formal semantics since the beginning (Frege 1892; Tarski 1935, 1956). Various phenomena, however, may prevent a sentence from being definitely true or false: vagueness, presupposition, oddness, or suboptimal phrasing of one kind or another can all give rise to truth-value judgments that are at least uncomfortable, if not outright impossible. We will describe two methods to identify truth-value gaps empirically. We will do so with a focus on a phenomenon known as *homogeneity*, which arises with plural definite descriptions. Our primary concern is its projection behaviour from the scope of quantifiers, which may be seen as an analogue of the well-known projection problem for presuppositions (Karttunen and Peters 1979, Heim 1983, and many others since).

1.1 Homogeneity

Sentences with definite plurals and distributive predicates have truth-conditions that are (roughly; see Sect. 1.3) the same as those of universally quantified sentences. Thus, the sentences in (1) are true of the same situations.

(1) a. Every book is written in Dutch.

b. The books are written in Dutch.

However, the observation has long been around in the theoretical literature (since Fodor 1970) that the negation of such sentences behaves differently. The negation of a universally quantified sentence is true whenever the non-negated sentence is not true; thus, the truth-conditions of (2a) are complementary to those of (1a).[1] This is not so for (2b): this sentence is only true if no book is in Dutch. In a situation where exactly half of the books are in Dutch, neither (1b) nor its negation (2b) are true.

(2) a. Every book isn't written in Dutch.
b. The books aren't written in Dutch.

If we identify the truth-conditions of the negation of a sentence S with the falsity-conditions of S, then we find that there are situations where a sentence with a plural definite is neither true nor false. The set of such situations is called its *extension gap* or *truth-value gap*.

The phenomenon just described is known as *homogeneity* (sometimes also as *polarity*) and has been discussed bySchwarzschild (1994), Löbner (1987, 2000), Breheny (2005), Gajewski (2005), Büring and Križ (2013), Spector (2013), and Magri (2014). The name is motivated by the fact that for distributive predicates, it can be formulated as follows: predication of a distributive predicate P of a plurality X only has a truth value if either all or none of the members of X satisfy P. In other words, these sentences have a truth-value gap if the members of X are not homogeneous with respect to the P property.

In this paper, we will develop and deploy methods of detecting and characterizing this truth-value gap.

1.2 Projection of homogeneity

Formally, the situation is similar to what happens in the realm of presupposition or vagueness: truth and falsity do not exhaust the possibilities. A presuppositional sentence may suffer from presupposition failure and be merely inappropriate and impossible to evaluate as being true or false (as in the famous *The king of France is bald*), while a sentence containing a vague predicate may fall in a gray area between truth and falsity (as in *My 180cm tall brother is tall*). The literature on these topics contain numerous proposals with respect to the nature of these "gappy" phenomena (see in particular Zehr 2014b and Spector 2015 for recent attempts at a joint description of these two types). One source of evidence has been the projection behaviour: if we can understand how operations like negation and quantification influence the distribution of truth-value gaps, we should be able to learn something about their nature. For instance, if we can show that a given operator influences differently the gaps resulting from two different phenomena, then it provides a strong argument for viewing these phenomena as conceptually distinct and treating them differentially.

More recently, the "projection" question has been raised with respect to truth-value gaps due to homogeneity: what happens to the truth-value gap when a definite plural is embedded in a complex linguistic environment, for instance under a quantifier? The traditional view of homogeneity as a presupposition (Schwarzschild 1994, Gajewski 2005) implies certain predictions about this question. But alternative views are now also being advanced: based on ideas in Spector (2013), Križ and Spector (2015) develop a theory related to the supervaluations approach to vagueness, and Magri (2014) presents a story based on scalar implicatures/exhaustification.

[1] Note that we are talking about the reading where negation has scope over the universal, which is probably the only one available for (2a) in English. An equivalent sentence would be *Not every book is written in Dutch.*

We will introduce more details about these theories in Sect. 6. For the moment, it is sufficient to note that what is crucial for all of these theories is how homogeneity effects play out when triggered in environments of different monotonicity. Accordingly, our empirical explorations will cover upward-entailing, downward-entailing, and non-monotonic environments.

1.3 Non-maximality

The empirical picture is complicated by a second phenomenon that has been noted in connection with plural predication. As a matter of fact, definite plurals are frequently — and felicitously — used in situations in which homogeneity is violated~~, and felicitously~~. Since Brisson (1998), this is known as *non-maximality*. For example, Lasersohn (1999) presents the example of (3), which certainly does not require every single one of the townspeople to be asleep in order to adequately characterize a situation.

(3) The townspeople are asleep.

The permissible degree of non-maximality seems to be highly context-dependent, and it has been argued that non-maximality can even, in some cases, result in effectively existential readings. Inspired by Krifka (1996), Malamud (2012) gives the following example:

(4) Context: Mary is going on a road trip with a friend. A few minutes into the ride, she exclaims:
Oh my, we have to go back — I left the windows open!

For this utterance to be a felicitous description of the situation, it seems to be sufficient that a single window has been left open.

Non-maximality thus refers to the possibility of using a sentence in a gap situation felicitously as if the sentence were simply true, rather than without a truth value. We may thus say that non-maximality has the effect of narrowing the observable extension gap, possibly to the point of closing it altogether in some cases. In other words, non-maximality can effectively fill in the gap with *truth*.[2]

1.4 Finding the third value

1.4.1 Experimental choices and past work

The experimental determination of the truth value of sentences that are capable of assuming more than two truth values (or failing to have any truth value at all) — whether due to presuppositions, vagueness, or homogeneity — is not a straightforward affair. Everyday usage only supplies us with the words *true* and *false*, and there is no third term that is entirely on a par with those and which could be unambiguously associated with the status we are after. Three experimental possibilities come to mind for diagnosing a third truth value. First, one could ask for ternary judgments. Despite the problem mentioned above, one could ask whether a sentence in a given situation is true, false or something else (the third option can be expressed in different ways). Second, one may collect standard truth-value judgments for both positive sentences and their corresponding negations, thus identifying a gap as a discrepancy between the two cases. Third, one may try to collect true and false judgments along with some independent behavioral measure, ranging from processing times or sophisticated ERP responses.

[2]Note that for sentences with a universal determiner instead of the definite (e.g., *I left all the windows open!*), the gap is filled in with *falsity* (at least if the environment is upward-entailing).

Note that we do not want to commit to the frequently made assumption that homogeneity is a presupposition. There may be different ways in which a sentence can fail to be either true or false, but for principled reasons, all of them pose the same problem to the experimentalist. As to the status of homogeneity vis-à-vis vagueness, see for instance Burnett (2013) for an analysis that relates homogeneous plural predication and vagueness in gradable adjectives.

In the case of presuppositional sentences, all three of these experimental options have indeed been tried, as we will now exemplify.

- Abrusán and Szendrői (2013) asked people to judge sentences with non-referring definites (e. g., "The king of France is bald" and "The king of France is not married to Carla Bruni") and gave them three options: *true*, *false*, and *can't say*. They obtained interesting results, but participants did not make much use of the third option. We suspect that this is due to the epistemic nature of *can't say*. The second method we we describe in the present paper can be seen as a variant of this approach.

- Chemla and George (2015), investigating a suspected presupposition of *agree* with a *wh*-complement, asked speakers separately for *true/false* judgments about positive and negated sentences. This method is only applicable to a phenomenon whose interaction with negation is known. Since there is no question that presuppositions project through negation, it is appropriate to their question, but we should avoid making the same assumption for homogeneity. Furthermore, the method could not be applied to sentences that do not have natural syntactic negations, like (5a), and therefore does not generalize to a broader investigation of the projection question.

  (5) a. Exactly 2 teachers love their students.
  b. ??Not exactly 2 teachers love their students.

- Schwarz (2013b) took a different avenue. He presented subjects with pictures and sentences that involved definite descriptions (e. g., *The circle on the left was green*), which sometimes failed to refer, and asked them to judge the sentences true or false. What he found is that a sentence is judged false both if the referent of the definite does not fulfill the predicate of the sentence and if the definite fails to refer altogether, but that response times are longer in the latter case. Thus, the theoretician's truth values 1, 0, and "presupposition failure!" translated to the judgments "true," "false with short response time," "false with long response time."[3]

  While Schwarz's results offer a distinctive characterization of judgments in truth-value gap situations (in that such judgments take significantly longer than regular *false* judgments), reversing the direction of inference and using response times to infer the status of the judgment would be complicated: it would require some objective fast/slow clustering method. For this reason, the methodology is well-suited to characterizing truth-value gaps, but less so to *detecting* them.

Truth-value gaps due to vagueness have been subject to experimental investigation as well.

[3]In a different, highly relevant experiment, Schwarz (2013a) collected judgment times for sentences with definite plurals, such as "The circles are green". Schwarz found that in non-homogenous situations (i. e., if some, but not all of the circles were green), non-maximal (true) judgments took significantly longer than maximal (false) judgments. For these results to be comparable to the results of Schwarz (2013b) discussed here (i. e., *false* judgments in a case of presupposition failure ~~begin~~being a slower version of *false* judgments in the face of plain falsity), one would need to know whether *false* judgments in non-homogeneous situations are slower than *false* judgments in cases of falsity.

- Alxatib and Pelletier (2011) gave subjects the three options *true*, *false*, and *can't tell*. Like Abrusán and Szendrői (2013), they found that the third option was largely ignored by the subjects. What they also found was that subjects were inclined to judge both a sentence and its negation *false* with respect to a borderline case (in line with Schwarz's 2013b results, see also Egré, de Gardelle, & Ripley 2013), but as discussed above, this effect could not be harnessed for our purposes.

- Ripley (2011), Sauerland (2011), and Zehr (2014a) asked their subjects to provide judgments on a so-called Likert-scale; that is, they offered multiple, ordered choices with extreme values labeled as *truth* and *falsity*. A value in the middle of the scale is slightly underspecified but could in principle indicate a gap judgment.

- Serchuk, Hargreaves, and Zach (2011) is situated at the intersection of the two paradigms described above: participants were presented with a range of labeled choices, including: *true*, *false*, and *not true, but also not false*, as well as some additional choices not directly relevant here. Contrary to the studies by Alxatib and Pelletier (2011) and Abrusán and Szendrői (2013), subjects did make use of the third option, which makes the approach a good candidate for a method for detecting gap judgments as well as vagueness judgments.

  Several aspects of the paradigm may be advantageously simplified for our purposes. First, the additional response choices (those not referring directly to the truth value) may simply be dropped. Second, note that in this experiment, the context for a given sentence was presented through a vignette which emphasized the borderline aspect of the situation. So a sentence such as "Susan is rich" was typically preceded by a vignette such as the following:

  > "Imagine that on the spectrum of rich women, Susan is somewhere between women who are clearly rich and women who are clearly non-rich. We are interested in your opinion about the status of the following [...] sentences."

  This vignette orients attention quite explicitly towards the third response option. Yet for more complex examples, which are needed to address the projection question, it is not obvious that the borderline nature/gappiness of the example can be communicated in the same way without using complex descriptions or revealing the purposes of our manipulation too openly. Instead, we will use pictures, which hopefully make an efficient replacement for textual descriptions and do not unduly call attention to the gap aspect of a situation.

1.4.2 The current contribution

We present two ways of diagnosing the failure of a sentence to be either true or false. Our first method relies on an independent evaluation of truth and falsity. A gap is found if a sentence is judged to be both *not true* (rather than *true*) and, on a different occasion, *not false* (rather than *false*). We introduce this methodology with Experiment A0 in Sect. 2 and apply it to more complex sentences in Sect. 3 (Experiments A1-3). We then replicate our findings with our second method (Sect. 4, Experiments B1-3): subjects are given three response options, as in previous studies, but these options are labelled differently (*completely true*, *completely false*, *neither*). We thus show that with the right choice of options, this is a viable method, too. In Sect. 5, we extend the latter methodology with small modifications and use it to obtain a more complete picture of the empirical situation (Experiments C1-4).

As already mentioned, at the theoretical level there are three main approaches to homogeneity (based on scalar implicatures, presuppositions, and supervaluations), and several ways

of fleshing each of them out to derive homogeneity projection. In Sect. 6, we will explore those, point out how our experimental results constrain them, and suggests further empirical tests. For now, it is sufficient to say that monotonicity plays a critical role in current theories, and that therefore we will investigate a range of upward-monotonic, downward-monotonic, and non-monotonic environments.

What we find is, in a nutshell, the following: the extension gap exists in simple sentences as common wisdom has it, and it projects through sentential negation. It also projects from the upward-entailing scope of the quantifier *all* as expected on most theories (neither universally nor existentially). It projects from under the non-monotonic quantifier *exactly*, albeit in a restricted way. Finally, we find that there is also homogeneity projection from the scope of *no*, but it is markedly weaker than from under *all*.

1.4.3 Roadmap of the current experiments

Our empirical investigation consisted of a series of 10 experiments: we provide here a roadmap to navigate between them. We had three batches of experiments, A, B and C, which vary in terms of tasks and types of stimuli; Table 1. We tested gaps of different kinds in 5 embedding contexts, as shown in Table 2, where framed cells indicate where evidence for a gap was found.

In all experiments, we tested potential gap conditions, but we also included clear-cut cases which served as reference points: either the same sentences in situations where it was clearly true or false, or the same situation accompanied by a sentence in which the definite description modified by *all* (which supposedly annihilates the gap), or clearly different sentences with quantified phrases which should not give rise to homogeneity effects (e.g., *most squares*, *few squares*).

	Stimuli	Task	Response options
Batch A	abstract (Fig. 1 or 7)	binary	{completely true / not completely true} and (different group): {completely false / not completely false}
Batch B	abstract (Fig. 11)	ternary	{completely true / completely false / neither} (for everyone)
Batch C	concrete (Fig. 15)	ternary	{completely true / completely false / neither} (for everyone)

Table 1: Overview of methods used, with type of stimuli and response options seen by a given participant. See Sect. 2.2 for discussion of labelling choices and response input method.

2 Experiment A0: developing method A for finding the gap

In this section, we present a method to detect gaps in simple environments. The paradigm we present in Sect. 2.1 consists in collecting *completely true* vs. *not completely true* judgments from a first group of participants and *completely false* vs. *not completely false* judgments from another group of participants. We then compare truth and falsity judgments, diagnosing a gap if a sentence is judged neither *completely true* nor *completely false*. In Sect. 2.2 we discussion certain seemingly minor aspects of the experimental design that turn out to be critical for robust results.

		Experiments											
E-∅	GAP	[A0]		[A1]	[B1]		[A2]	[B2]		[A3]	[B3]		
E-*neg*	GAP			[A1]	[B1]								
E-*all*	GAP						[A2]	[B2]	[C2]	[A3]	[B3]	[C3]	[C4]
E-*no*	GAP						A2	B2	[C2]				
E-*exactly*	GAP									[A3]	[B3]	[C3]	[C4]
	GAP?									A3	B3	C3	
	GAP??												[C4]

Table 2: Embedding contexts for gap testing. All conditions were multiplied along three dimensions in a factorial manner: (i) TRUE and FALSE situations were added to the critical GAP conditions, (ii) besides the target version of a sentence with a definite plural, a control version with a modifying *all* (supposedly removing homogeneity effects) was tested); (iii) control sentences with *most* and *few* as quantifiers were added. Frames mark situations in which evidence for a gap was found.

2.1 Experimental design and analysis

We independently assess truth and falsity for a sentence, from two groups of participants involved in a *completely true* vs. *not completely true* judgment task (Fig. 1a) and a *completely false* vs. *not completely false* task (Fig. 1b), respectively. A gap would be found if a sentence is judged *not true* by the first group of participants and *not false* by the other group, which situation corresponds precisely to what an extension gap is: neither truth nor falsity.

2.1.1 Participants

We recruited two groups of 20 participants via Amazon's Mechanical Turk, who were paid a fixed sum for their participation in the experiment. One participant took part in both groups and was excluded from further consideration, leaving us with 38 participants overall, all of whom were native speakers of English. There were 13 females and 25 males, aged between 21 and 60 years.

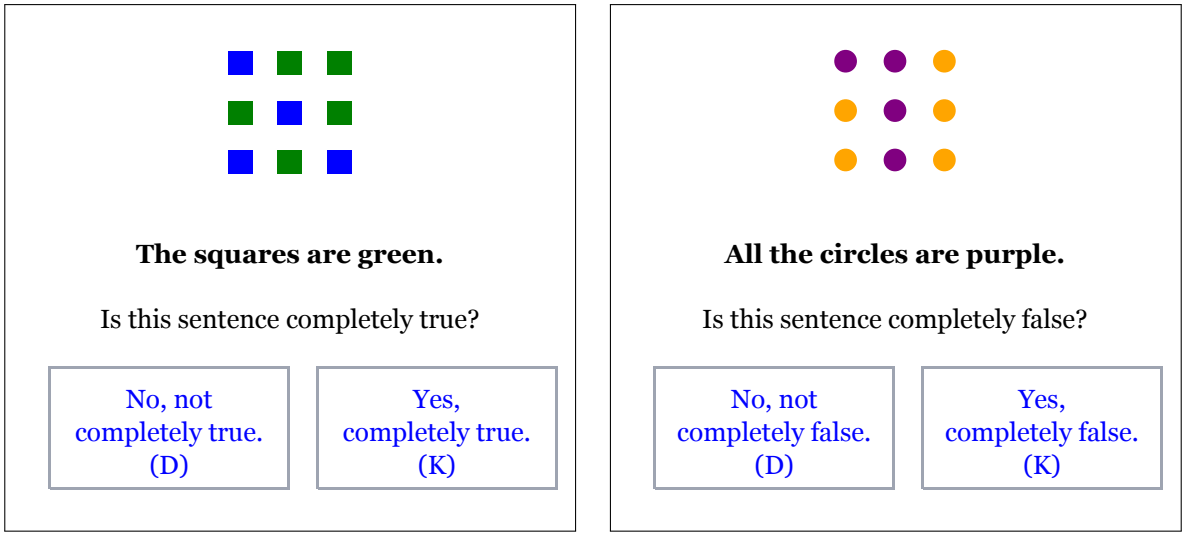
2.1.2 Materials

Each item consisted of a sentence and a display, as exemplified in Fig. 1.

Sentences Participants were asked to judge sentences of four kinds, as schematized in ((6)). The set of possible shapes, both in the sentence and in the display, was: triangles, circles, squares, and stars. Possible colors were: red, blue, green, orange, and purple. We refer to these sentence types by their determiners, as follows: **the** is our target sentence, **all** is a control, and **few** and **most** are fillers.

(6) a. *Target sentences*
The [shapes] are [color].
Example: The triangles are blue.

b. *Control sentences*
All the [shapes] are [color].
Example: All the circles are orange.



(a) Target (**the**) sentence in the *completely true* vs. *not completely true* task

(b) Control (**all**) sentence in the *completely false* vs. *not completely false* task

Figure 1: Examples of items from Exp. A0.

c. *Fillers*
Few [shapes] are [color].
Most [shapes] are [color].
Examples: Few squares are red.
Most stars are purple.

Displays Participants were asked to judge each sentence against a display that contained 9 of the shapes mentioned in the sentence. The displays can be classified into 10 different kinds, depending on how many shapes were of the target color mentioned in the sentence. Let us refer to these as 'target-color shapes'. A display could thus contain 0, 1, ..., or 9 target-color shapes. The remaining of the 9 shapes were all of the same color from the same group of colors given above.

Combinations The **the**-sentences and the **all**-sentences (i. e., targets and controls, respectively) each appeared with each kind of display 4 times, once with each of 4 kinds of shapes: triangles, circles, squares, and stars. Thus, every participant saw 40 target sentences and 40 control sentences. We then obtain three conditions: (1) the FALSE condition, with 0 target-color shapes; (2) the TRUE condition, with 9 target-color shapes; and (3) the GAP condition, with both target and non-target-color shapes, which violates the homogeneity assumption. Critically, a **the**-sentence should show a gap behavior for this last condition (corresponding to Diagnostic 2), while its **all**-counterpart should be false (Diagnostic 1).

Fillers with **few**- and **most**-sentences were used to make up 60 items each. The sentences were paired with all possible displays (as defined by the number of shapes of the target color), resulting in 40 items for each filler determiner, to which we added 20 items per filler determiner that utilitized displays with 1 to 5 target-color items for **few** and with 4 to 9 target-color items for **most** in order to balance the ratio of true and false descriptions of displays.

Presentation The items were presented in random order, except for an additional group of 4 simple introductory items, which were presented at the beginning of the experiment to acquaint

participants with the task before they would be presented with critical items.

2.1.3 Procedure

Participants were first assigned to a particular group, either **T(ruth)** or **F(alsity)**, and were given instructions accordingly. **T**-participants were asked to judge whether the sentences were *completely true* or *not completely true* of the display, as in Fig. 1a. **F**-participants were asked to judge whether the sentences were *completely false* or *not completely false*, as in Fig. 1b. Participants entered their responses by either clicking their choice with the mouse or pressing a key ('D' and 'K', respectively).

At the end of the experiment, participants completed a questionnaire gathering demographic information such as native language, age, and sex. Participants were also asked whether they were color-blind. Across all experiments presented here, 28 out of 674 participants reported to be color-blind and including or excluding their responses in the analyses did not affect the results in any way. (We might note here that our choice of colors aimed at minimizing the difficulty of perceiving a relevant contrast.) The analyses we present below, for all experiments, include these participants.

2.1.4 Data treatment and analyses

We excluded from the analysis three participants whose accuracy was less than 75% on unambiguous filler items. Our results are thus based on responses from 16 **T**-participants and 19 **F**-participants.

The statistical tests presented below are based on logit mixed-models with a maximal structure for the random effects subject and item; we report the χ^2 and p-values obtained by performing likelihood ratio tests in which the deviance of a model containing the fixed effect of interest was compared to another model without it, but otherwise identical in random effect structure (as recommended in Barr, Levy, Scheepers, & Tily 2013).[4] Error bars in the graphs correspond to standard error to the mean per participant.[5]

2.1.5 Result presentation

For this first experiment, we will represent the results in three figures (Figs. 2, 3, and 4) containing largely overlapping information albeit from different perspectives.

Figure 2 provides results for each combination of sentence and display, before aggregation by type of display. We use it to make one informal but important remark. The target **the**-sentences in the GAP condition (1 to 8 target-color shapes out of 9 shapes) show a discrepancy between the two tasks, as revealed by a high number of *not completely true* ratings coexisting

[4] In some cases, the logit models did not produce meaningful results: some effects that *look* to be the strongest were not recognized as such by our statistical tests. We gathered that this was due to ceiling effects that impact the log-estimation behind logit models. For this reason, in cases where we did not observe a significant result, we went on to run the same model but on a modified data set whereby 5% of the responses would have been changed randomly. This manipulation was designed to get rid of ceiling effects, but be ~~highly conservative~~conservative (it is unlikely for noise to lead to the appearance of a new effect). One may worry that the threshold of significance we consider should be altered accordingly (as in more standard multiple comparison situations). But note that here the problem is well-identified and pertains to the analysis not being applicable (as in situations in which there is a failure of a statistical model to converge or to meet assumptions for it to be applicable). In any event, we typically observe p-values which are far from the threshold, so standard corrections would not affect our conclusions.

[5] Graphs and statistics were created using the R programming language (R Core Team 2014) with the libraries `ggplot2` (Wickham 2009), `lme4` (Bates, Maechler, Bolker, and Walker 2014), `reshape2` (Wickham 2007), and `plyr` (Wickham 2011).

with a high number of *not completely false* ratings. This pattern is evidence of a gap, and it can be observed across the whole potential gap — that is, regardless of the precise number of target-color shapes, as long as it is not 0 or 9.

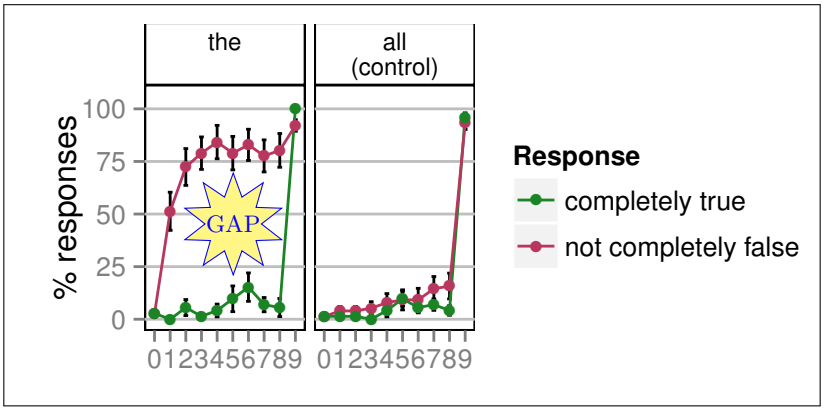


Figure 2: Main results from Exp. A0 before aggregation of displays into TRUE, FALSE, and GAP conditions. For the target **the**-sentence with 1 to 8 target-color shapes, the distance between the red line (revealing a high rate of *not completely false* responses) and the green line (low rate of *completely true* responses, i.e. high rate of *not completely true* responses) reveals the gap.

Figures 3 and 4 provide the same results for the target and control sentences in the FALSE (0 target-color shape), the TRUE (9 target-color shapes) and the *aggregated* GAP condition (1 to 8 target-color shapes). Figure 3 separates the results of the **T**-participants from those of the **F**-participants. Figure 4 is a different visual arrangement of the same data, stacking on top of each other the average of *completely true* responses from the **T**-participants and the average of *completely false* responses from the **F**-participants.

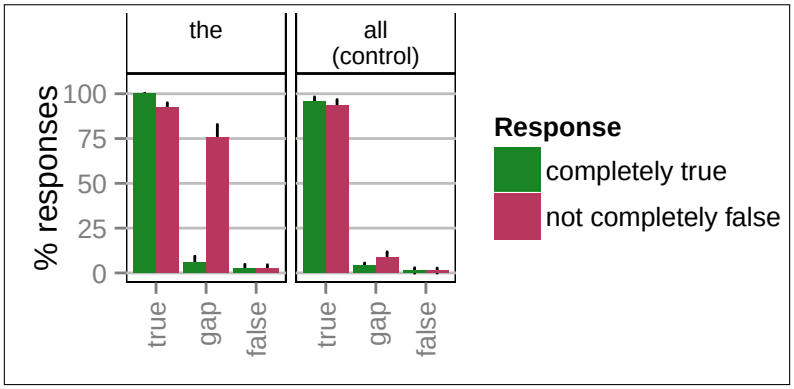


Figure 3: Main results from Exp. A0 for **T**- and **F**-participants shown for FALSE (0 target-color shape), TRUE (9 target-color shapes), and GAP (1 to 8 target-color shapes) conditions.

Before moving on to more important questions, let us observe that participants responded as expected to the control **all**-sentences ((6)b) in all three conditions as well as to the target sentences in both control TRUE and FALSE conditions (accuracy is above 91% in all cases for both **T**-participants and **F**-participants).

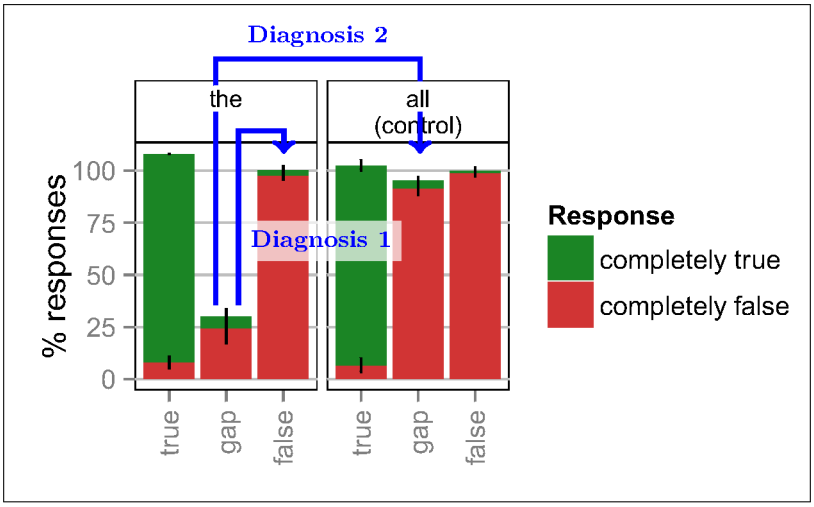


Figure 4: Main results from Exp. A0 for FALSE (0 target-color shape), TRUE (9 target-color shapes) and GAP (1 to 8 target-color shapes) conditions. The *completely true* responses from **T**-participants and the *completely false* responses from **F**-participants are stacked on top of each other (with error bars at the top of each bar), so that the gap in the GAP condition for **the**-sentences is revealed by the failure to reach the ceiling score of 100% when both truth and falsity judgments are added together.

2.1.6 Main results: three diagnostics of a gap

We are interested in how target sentences behave in the GAP condition, both compared to the control sentences (which do not show a gap and are judged false) and compared to the FALSE condition (which provides a baseline to distinguish a gap from falsity). In a nutshell: a gap is found if true and false responses do not exhaust the possibilities — that is, if the bar in the GAP condition for the target sentence does not fill in the whole space (or, in comparative terms, does not reach as high as the corresponding bar in the FALSE control condition or for the **all** control sentence).

Concretely, we will look at three (non-independent) diagnostics for the existence of a gap. At the practical/statistical level, these diagnostics are best captured when considering the data as organized in Fig. 3, but Fig. 4 provides a more convenient visual representation of them.

Diagnostic 1: Comparison between 'gap' and 'false' conditions. Evidence for a gap would be found if target sentences behave differently in the GAP condition than in the FALSE condition. Visually, this is evidenced by the fact that for the target sentence the stacked bar from Fig. 4 is lower in the GAP condition than in the FALSE condition. This demonstrates that *completely false* and *completely true* judgments do not add up to 1 as in a regular FALSE situation.

Figure 4 is convenient to assess this claim visually, but in this form, there is no obvious statistical test for it.[6] Taking the perspective of Fig. 3, however, one can see that it is equivalent to check whether there is a statistical interaction between conditions (FALSE vs. GAP) and true-corresponding responses in the two tasks (comparing *completely true* vs. *not completely false*). If there is a gap, there should be more *not completely false* responses than *completely true*

[6]The reason for this is that the bars that are stacked on top of each other are based on responses from different participants, so that although the resulting total height is a useful visual guide, it does not make sense to add bars up pairwise, for instance (we could still add them based on items however, but for now we will use another technique to assess the same point.

responses for the target sentences, but only for the GAP condition (the FALSE condition serves as a baseline). In our data, this interaction is indeed statistically significant: $\beta = 5.6, \chi^2 = 347, p < 10^{-15}$ (see Sect. 2.1.4 for details about the underlying statistical model).

Diagnostic 2: Comparison between the- and all-sentences. Evidence for a gap would also be found if target **the**-sentences behave differently than the control **all**-sentences in the GAP condition. Visually, this is evidenced by the fact that in the GAP condition the stacked bar from Fig. 4 is lower for the target sentences than for the control sentences.

Technically, this amounts to asking whether there is a statistical interaction between sentences (**the** vs. **all**) and true-corresponding responses in the two tasks (comparing *completely true* vs. *not completely false*). If there is a gap, there should be more *not completely false* responses than *completely true* responses in the GAP condition, but only for the **the**-sentences. This interaction is indeed statistically significant: $\beta = 6.3, \chi^2 = 623, p < 10^{-15}$.

Diagnostic 3: Combined comparison with all-sentences and 'false' conditions. The two diagnostics above can be combined into a higher order type of evidence. A gap would be found if the difference between conditions FALSE and GAP is different for target **the**- and control **all**-sentences. Technically, the full interaction between sentences (**the** vs. **all**) and conditions (GAP vs. FALSE), on the responses in the two tasks (comparing true corresponding responses: *completely true* and *not completely false*) is statistically significant: $\beta = 5.25, \chi^2 = 6.15, p = .013$.[7]

2.2 Discussion: Fine-tuning the methodology

We proposed a methodology to assess truth and falsity independently. We used it to obtain three (non-independent) diagnostics which indicate that non-homogeneity creates a truth-value gap for definite descriptions. The first Diagnostic shows that this gap behavior is different from falsity, the second Diagnostic shows that it is not observed similarly for universal quantifiers, and the third Diagnostic is a combination of the first two.

We would like to highlight two critical aspects of our method.

- First, the presence of the adverb *completely* is essential. In a pilot experiment which was identical to Experiment A0 except for the absence of *completely* in the response options, we could not observe a clear gap, as seen in Figure. 5. We suspect that this is mostly due to a weak reinterpretation of *false* as *not true*, which subjects would not so readily apply to *completely false*. Furthermore, the presence of *completely* discourages non-maximal interpretations (as can be seen in the lower percentage of *completely true* responses in comparison with plain *true* responses), thus counteracting non-maximality's effect of narrowing the observable gap (see Sect. 1.3).

- Second, there remained a certain tendency for participants to assimilate the task to a *true* vs. *false* task, rather than a *true* vs. *not true* task. We formed this impression based on our own feeling about the way we did the task and reports from informants who commented on early versions. We noted that this tendency could be increased by the possibility to ignore the response labels once they have mastered the mapping between responses and keys. For this reason, in all subsequent experiments participants had to enter their responses by *clicking* on the appropriate response choice. Our reasoning was that this would lead

[7] We faced several failures to converge in this case. We thus used a simpler random effect structure with slopes for Participant and Item only for the sentence factor (**the** vs. **all**), although the other random effects were correlated with it in the more complex models we obtained (before plain convergence).

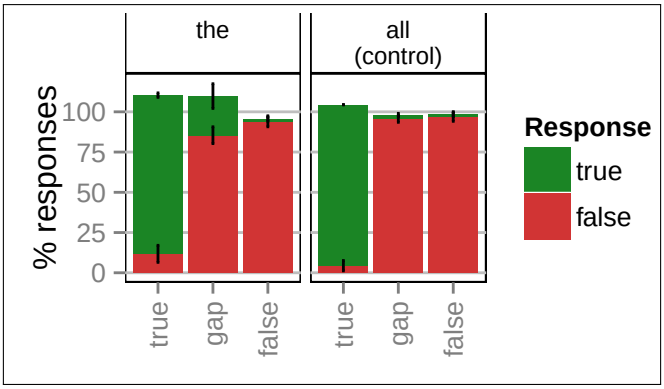


Figure 5: Results from a pilot study which did not use the adverb *completely* in the response choices. Specifics for this figure are otherwise comparable to Fig. 4.

them to see and possibly read the response label on each and every occasion. Our hope was to attract their attention to the fact that the *not (completely) true* option was not similar to a *(completely) false* option.

2.3 Summary

As laid out in this section, we developed an experimental two-response paradigm that can detect a truth-value gap. In Sect. 3, we will apply this method to the projection problem of homogeneity. In Sect. 4 and 5, we will discuss a method based on a three-response choice. As we will see, the parallel use of all of these methods offers an opportunity for replication of results and a cross-validation of methods.

3 Experiments A1-A3: Application to projection

With this methodology tested and in place, we set out to apply it to the projection of homogeneity from upward-entailing, downward-entailing, and non-monotonic contexts. In order to test for the presence of a gap, we need at least some idea of where to look for itthat is, some plausible candidate situations that should be gaps for one sentence or another. A more detailed discussion of theoretical considerations will be deferred until Sect. 6, but let us state a simple guiding principle here. Our most promising theories predict something not necessarily identical to, but close to, the following: a sentence S with a definite plural has a gap in a given situation if a variant sentence S', where the definite description in S is replaced with an existential is true in that situation, while the variant S'' where the definite is replaced with a universal quantifier is false.

For example, (7) would have a gap in a situation where (7a) is true while (7b) is false.

(7) All the teachers like their students.
   a. All the teachers like some of their students.
   b. All the teachers like all of their students.

Unless otherwise noted, all our gap conditions were chosen on the basis of this rule.

3.1 Experiment A1: Negation

3.1.1 Participants

We recruited two groups of 30 participants via Amazon's Mechanical Turk. Two workers did not do the task and one participant was excluded from the analyses for not being a native speaker of English, resulting in two final groups of 30 and 27 participants, respectively. There were 22 females and 35 males, aged between 19 and 60.

3.1.2 Materials

Sentences We used the same sentences as in Experiment A0, but added negated versions of the target and control sentences. The sentences of interest thus varied along two factors, determiner (**the** vs. **all**) and **e**nvironments (**E-∅** and **E-neg**). Fillers were again formed with **few** and **most** (in positive versions only). Thus, all sentences were of one of the forms below.

(8) *Target sentences*

a. The [shapes] are [color]. (**the**, **E-∅**)
b. The [shapes] aren't [color]. (**the**, **E-neg**)

(9) *Control sentences*

a. All the [shapes] are [color]. (**all**, **E-∅**)
b. Not all the [shapes] are [color]. (**all**, **E-neg**)

(10) *Filler sentences*

a. Few [shapes] are [color]. (**few**, **E-∅**)
b. Most [shapes] are [color]. (**most**, **E-∅**)

Sentence type (9b) was chosen so as to make explicit that negation takes wide scope, since we were trying to test embedding under it. The preliminary assumption is, of course, that the definite description in (8b) also does not take distributive wide scope over negation.[8] See Sect. 6.1.3 for further discussion.

Displays We used the same arrays of colored shapes as in the previous experiment. Of the 9 shapes, n were of the target color mentioned in the sentence, where n assumed the values $0, 2, 4, 6, 8, 9$.

Combinations Every combination of determiner, environment, and number of target-color items was instantiated 4 times, once for each kind of shapes (triangles, circles, squares, and stars). This yielded 48 target items, 48 control items, and 48 filler items for a total of 144 items per participant.

Presentation As before, the items were presented in random order, except for an additional group of 4 simple introductory items, which were presented to participants at the beginning of the experiment.

3.1.3 Procedure

The procedure was the same as in Experiment A0 except that participants entered their response by clicking on the response labels with the mouse rather than using keys (see discussion in Sect. 2.2).

[8] Cf., e.g., Steedman (2012) for arguments that definite plurals do not take derived distributive scope.

3.1.4 Results

Figure 6 represents the results for positive sentences (as before) and the negative sentences (new), with controls containing a universal quantifier. Again, participants with less than 75% accuracy on uncontroversial filler items were excluded. This affected 5 participants. As before, responses in uncontroversial situations are accurate ($M > 89\%$ except for the TRUE conditions for the **all**-controls, which seemed particularly difficult). We also observe that the three gap diagnostics apply positively to the two types of sentences: see Table 3 for details. These results therefore replicate the previous finding (Sect. 2) of a gap for unembedded definite descriptions and extend it to negative sentences.

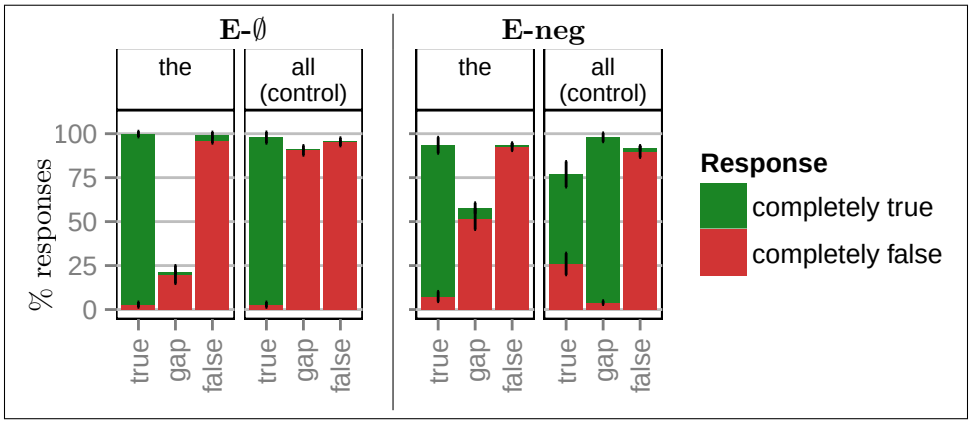


Figure 6: Main results from Exp. A1 (For plotting details see Fig. 4.)

	E-∅			**E-neg**		
	β	χ^2	p	β	χ^2	p
Diag. 1	-7.8	13.2	.0003	-3.8	9.80	.002
Diag. 2	-5.1	9.79	.002	-4.4	19.8	10^{-5}
Diag. 3	-35	12.6	.0004	-4.9	11.7	.0007

Table 3: Experiment A1: Results of the statistical diagnostics of a gap for positive (**E-∅**) and negative (**E-neg**) sentences. All diagnostics yield significant results for **E-∅** and **E-neg**.[9]

3.1.5 Discussion

These results replicate the previous finding of a gap for unembedded definite descriptions (**E-∅**) and extend it to negative sentences (**E-neg**). We postpone the discussion of the consequences of this first projection result for existing theories to Sect. 6, and proceed to investigate further environments.

3.2 Experiment A2: Embedding under *all* and *no*

3.2.1 Participants

We recruited a total of 81 participants, separated into two groups, via Amazon's Mechanical Turk. Three participants were excluded for having completed a previous version of our experi-

[9] In looking at such tables, the p-values should be corrected using the Holm-Bonferroni method. Our interpretations are based on the output after such a correction is applied. In most cases, sufficiently many p-values are clear-cut such that this correction does not affect the significance of any individual result.

ment, and three more for being native speakers of a language other than English, leaving us with two groups of 37 and 38 participants, respectively. Of these, 35 were female and 39 male, with one participant identifying with neither category. They were aged between 21 and 64 years.

3.2.2 Materials

Sentences Participants were asked to judge sentences of the three following kinds, which vary with respect to how the key determiner is embedded. The **E-∅** kind of sentences are as before, without embedding, while **E-*all*** and **E-*no*** sentences introduce a level of embedding under *all* and *no*, respectively.

(11) **E-∅**: [determiner] [shapes] are [target color].

(12) **E-*all***: In all the cells, [determiner] symbols are [target color].

(13) **E-*no***: In none of the cells, [determiner] symbols are [target color].[10]

As before, the determiner was 'the' in the target conditions (**the**) and 'all the' in the controls (**all**), while **few** and **most** were again used for fillers. Some concrete examples for the new types of sentences are given in (14).

(14) a. In all the cells, the symbols are green.
b. In all the cells, all the symbols are blue.
c. In none of the cells, all the symbols are purple.
d. In none of the cells, most symbols are red.

Displays Besides the two factors discussed above (embedder and determiner), which concerned the form of the sentence, the third factor in our experiment was the truth value that the sentences was assumed to have with respect to the display it was paired with. Below, we describe only the TRUE, FALSE, and GAP displays used with the target and control sentences.

For sentences with no embedder (i. e. **E-∅**), the displays were like those in the previous experiments, containing only one array of 9 shapes. Here, the TRUE condition had all of the (relevant kinds of) shapes of the target color, the FALSE condition none, and the GAP condition some, but not all.

In order to add a layer for an embedding quantifier in **E-*all*** and **E-*no*** sentences, we used displays containing 4 arrays, one for each kind of shape, resulting in items like the one exemplified in Fig. 7. For the embedder **E-*all***, the TRUE condition had only target-color shapes everywhere. The FALSE condition had at least one cell with no target-color items. The possibilities that remain were classified as GAP. Analogously, for **E-*no***, there were no target-color shapes at all in the TRUE conditions, and at least one cell with all shapes in the target-color in the FALSE condition.

Combinations The target and control conditions with **E-∅** were instantiated 4 times (once for each kind of shapes), and those with **E-*no*** and **E-*all*** were instantiated 6 times. This resulted

[10] Due to our simply replacing *all* with *no* in the sentences, we accidentally ended up with sentences that were actually ungrammatical (thanks to Paul Égré, Irene Heim and Jeremy Kuhn for discussion). There should have been negative inversion, leading to this pattern:

(i) In none of the cells are [determiner] symbols [target color].

The sentences seem to be interpretable without difficulty, but of course this situation casts some doubts on the results. Experiment C2 corrects this issue, as well as some other concerns about the particular examples with *no* used here, cf. also footnote 14.

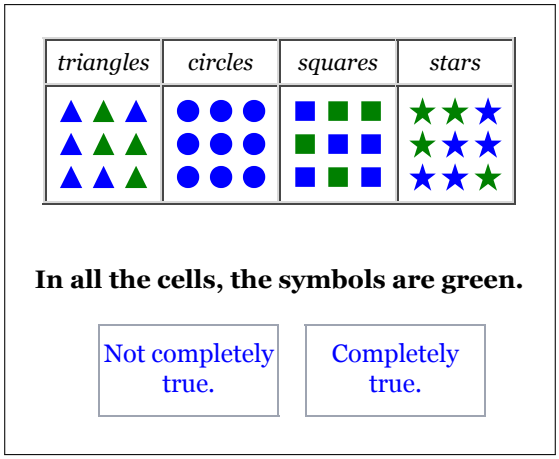


Figure 7: Example of an item involving quantification (**E-*all*+FALSE** example from Exp. A2).

in a total of 48 target items (embedded determiner **the**) and 48 control items (embedded determiner **all**). There were also 32 filler items for a total of 128 items per participant. More details about the instantiations of the conditions are given in Table 13 in the Appendix.

Presentation As before, the items were presented in random order, except for an additional group of 12 simple opening items.

3.2.3 Procedure

The procedure was the same as in Experiment A1.

3.2.4 Results

Due to a coding error, half of the **E-*no*+FALSE** conditions did not actually show situations where the sentences were false and were consequently excluded. Figure 8 provides the results in the same format as before, for both target and control sentences with all three embedders, after two participants were excluded for insufficient accuracy on uncontroversial fillers.

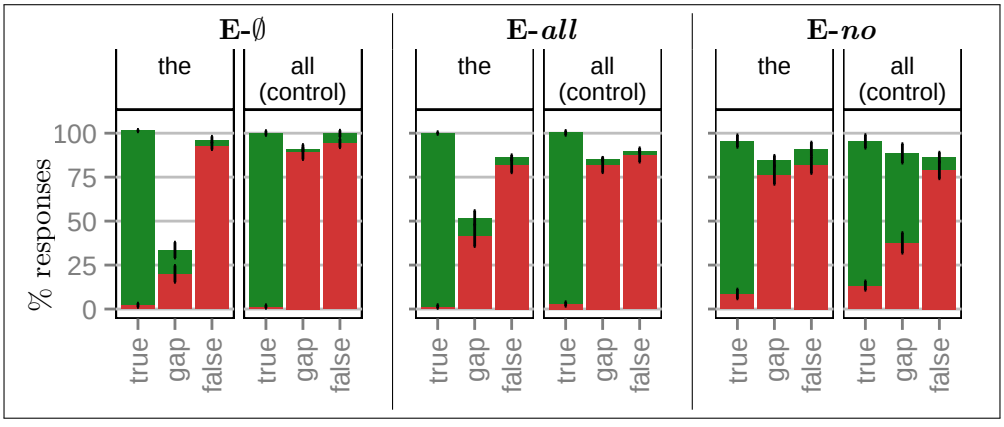


Figure 8: Main results from Exp. A2 (For plotting details see Fig. 4.)

The previous results for unembedded definites (**E-∅+the/all**) were again replicated. We also observed a gap for embedding under **E-*all***, but none for **E-*no***. The outcomes of our

	E-∅			**E-*all***			**E-no**		
	β	χ^2	p	β	χ^2	p	β	χ^2	p
Diag. 1	−4.9	28.9	10^{-7}	−3.8	6.97	.008	−.16	.074	.79
Diag. 2	−4.6	6.27	.012	−3.5	11.0	.0009	−1.4	2.17	.14
Diag. 3	−3.9	5.72	.017	−3.7	7.11	.008	−.69	.85	.36

Table 4: Experiment A2: Results of the statistical diagnostics of a gap for definite descriptions at various levels of embedding: **E-∅**, **E-*all***, and **E-*no***. All diagnostics yield significant results for **E-∅** and **E-all**, but not for **E-no**.

diagnostics are summarized in Table 4. In addition, we note a curiously high rate of *completely false* responses in the controls for the **E-*no***+GAP condition where a *completely true* response should have been given (cf. Table 13 in the Appendix). It seems that unlike the accidentally ungrammatical target sentences, such as (15), the accidentally ungrammatical control (16a) may actually be impossible to interpret (cf. also discussion of this matter in section 6.2). Compare this with the grammatical (16b), which clearly seems to have the intended bound reading for the definite.

(15) In none of the cells, the symbols are green.

(16) a. In none of the cells, all the symbols are green.
b. In none of the cells are all the symbols green.

If there were indeed participants who may have found (16a) uninterpretable, they may have given roughly chance-level responses for that reason. Let us note here that restricting the analysis to those participants who did respond with *completely true* in a higher percentage of cases in this condition did not in any way change the overall picture in the other conditions in any way.

3.3 Experiment A3: Embedding in non-monotonic environments

Finally, we went to see what happens if a definite plural is embedded under the non-monotonic quantifier *exactly*. Here, we tested two different kinds of conditions that different theories predict to be gaps.

3.3.1 Participants

We recruited a total of 80 participants via Amazon's Mechanical Turk. After exclusion of five non-native speakers of English, we were left with two groups of 40 and 35 participants, respectively, consisting of 37 females and 38 males, aged between 19 and 61 year.

3.3.2 Materials

Sentences The sentences used were identical to those in Experiment A2, except that the **E-*no*** sentences were replaced with **E-*exactly*** sentences, which had the form in (17).

(17) In exactly 2 of the 4 cells, [determiner] [shapes] are [target color].

Displays The displays were of the same kind as in the previous experiment (see Fig. 7). We created **E-*exactly***+TRUE conditions using displays with exactly two cells that had all shapes in the target color and the other two cells containing none. For **E-*exactly***+FALSE, either fewer

than two cells had any shapes in the target color, or more than two cells had all shapes in the target color.

We created two (potential) gap conditions. In **E-*exactly*+GAP**, exactly two cells had target-color shapes, but in at most one of these cells, all the shapes were of the target-color (if they had been in both, that would have constitued a TRUE situation). In **E-*exactly*+GAP?**, one cell had only target-color shapes, two cells had a number n of target-color shapes (n was 2, 3, 4, 5, 6, or 7), and the remaining cell had no target-color shapes.

Combinations In sum, we use the same items as in Experiment A2 plus 12 items for **E-*exactly*+GAP** (6 with the embedded determiner **all** and 6 with **the**), for a total of 198 items per participant. Again, a detailed table can be found in the Appendix.

Presentation As before, the items were presented in random order, except for an additional group of 12 simple opening items.

3.3.3 Procedure

The procedure was identical to that in Experiment A2.

3.3.4 Results

Our data are shown in the familiar format in Fig. 9. The outcomes of our diagnostics for gaps can be seen in Table 5. Results are clear in uncontroversial cases: the results from Experiment A2 for **E-*all*** and unembedded definites are replicated and we find evidence of a gap for **E-*exactly*** in the uncontroversial gap condition GAP. However, we did not find a statistically significant gap in the GAP? condition with **E-*exactly*.**

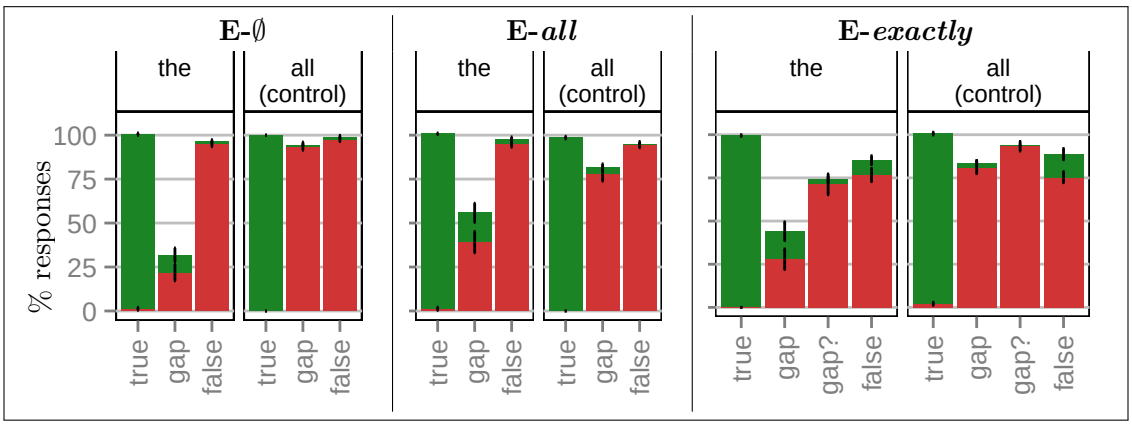


Figure 9: Main results from Exp. A3 (For plotting details see Figure 4.)

3.4 Intermediate Summary and Discussion

In the set of experiments reported in this section, we applied our methodology of assessing separately the truth and falsity of sentences to projection scenarios for homogeneity.

In Experiment A1, we found clear evidence for projection of homogeneity through negation; that is to say, negated plural predication displays the same extension gap as unnegated plural predication.

	E-∅			**E-*all***			**E-*exactly***					
	GAP			GAP			GAP			GAP?		
	β	χ^2	p	β	χ^2	p	β	χ^2	p	β	χ^2	p
Diag. 1	−4.5	24.7	10^{-6}	−2.5	10.5	.001	−9.1	10.7	.001	−3.4	5.27	.02
Diag. 2	−3.6	11.1	.0009	−1.14	4.57	.03	−6.4	12.7	.0004	−1.2	2.5	.11
Diag. 3	−3.9	10.3	.001	−1.5	3.91	.048	−4.7	5.45	.020	−0.46	0.48	.49

Table 5: Experiment A3: Results of the statistical diagnostics of a gap for definite descriptions at various levels of embedding: **E-∅**, **E-*all***, and **E-*exactly***, with the two gap conditions GAP and GAP? for the latter. All diagnostics yield significant results except for GAP? with **E-*exactly***.

In Experiment A2, we found projection of an extension gap through *all*, but no projection through *no*. For the theoretical significance of these findings, see Sect. 6.

In Experiment A3, we replicated the findings for *all* from Experiment A2 and furthermore tested for gaps with definite plurals in the scope of *exactly*. In condition GAP, a gap was found as expected, whereas the tests applied to condition GAP? revealed no robust gap. However, it is worth having a closer look at the graph (Figure 9) in the latter case. Two things might be noted in this graph: first, though it is not statistically significant, it is suggestive of a small gap for **E-*exactly***+GAP? (in line with the result of Diagnostic 1, see Table 5); and second, the number of *true* responses for FALSE is surprisingly high for both target sentences and controls. We interpret these observations in detail in the following two paragraphs, but let us give a summary before going into the specifics: there is overall strong evidence against the presence of a GAP? gap. We find little evidence of one to begin with, but even that small evidence can be explained away as a GAP kind of gap (together with the presence of a surprising but detectable *at least*-reading).

Let us now see how this plays out in detail. Looking at individuals items in the FALSE condition as presented in Figure 10, we can see a reason for the latter observation. We observe that there is a higher tendency to count the *exactly 2* sentence as true when *more than 2* applies (the three items coded as 9990, 9994, and 9999, where each digit indicates the number of target-color shapes in one of the four cells) than when *less than 2* does (0000, 09000, 4000). This pattern makes sense if we assume that *exactly 2* has a non-salient reading on which it ends up equivalent to *at least 2*. Evidence for such an *at least*-reading has been found by Marty, Chemla, and Spector (2014) for another non-monotonic modified numeral phrase, *between 3 and 5*. The proposed mechanism for how it arises is that the sentence is interpreted as asserting that "there is a group which consists of exactly 2 ... such that ..." For distributive predicates, this is,of course also the case if there is furthermore a group that consists of more than exactly 2 members. This *at least*-reading is true for exactly those three items for which we find a heightened rate of *true* responses.

Looking now at the GAP? condition, we find that the items in this condition should be gap cases for the *at least*-reading by the same reasoning by which GAP is a gap for the proper *exactly*-reading. All the items have one cell with all 9 shapes in the target color, and two more cells with some but not all shapes in the target color. Thus, the *at least*-reading is true if the definite is replaced with *some*, but false if it is replaced with *all*. We conclude, therefore, that to the small extent to which gap judgments are observed in the GAP? condition, they can be explained by the presence of an *at least*-reading and are not due to GAP? being a gap on the *exactly*-reading.

In the next section, we will show that all of these results can be reproduced by presenting a single group of participants with three response options, provided the options are of the right kind.

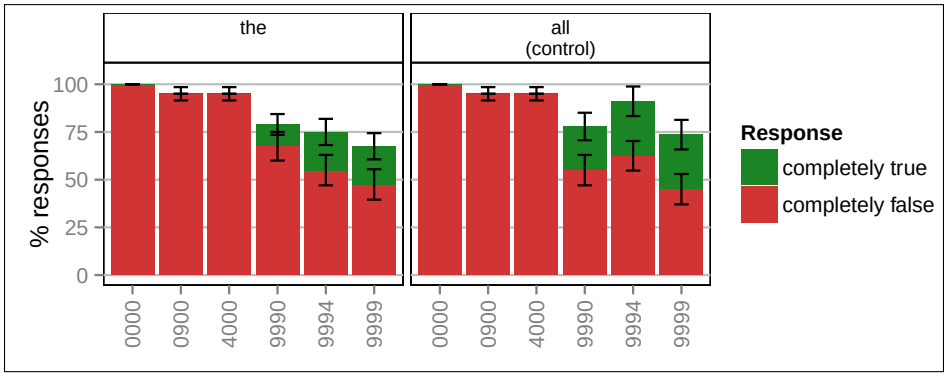


Figure 10: Responses split up per item in the **E-*exactly*+FALSE** condition. The code indicates the number of target-color shapes in each cell (e.g., '0123' represents an item with 0 target-color shapes in the first cell, 1 in the second, 2 in the third and 3 target-color shapes in the fourth cell).

4 Experiments B1-B3: three response choice

We present here an alternative method to gather gap judgments, which we used to replicate the results obtained in Experiments A2 and A3. This alternative method uses the same materials but with only one group of participants. The participants were given the response options *completely false*, *neither*, and *completely true*. This set of three responses, although simple, has, to our knowledge, never been used in comparable experiments. Our precise choice of response labels here was guided by our previous exploration of the possibilities (see the literature review and Sect. 2.2 above). A sample item is shown in Figure 11. In all other respects, our B-experiments were clones of our A-experiments.

Our three diagnostics translate to this new paradigm by directly comparing the rate of *neither*-responses in different conditions. In fact, it is striking that the result figures look exactly like the ones from our previous experiments, except that the *neither* responses fill in the space left empty in the previous, binary paradigm. The new versions of the diagnostics are thus as follows: diagnostic 1 compares the rates of *neither*-responses in GAP and FALSE conditions for **the**-sentences, diagnostic 2 compares rates of *neither*-responses in GAP conditions for **the**- and **all**-sentences, and diagnostic 3 investigates the interaction between conditions (GAP vs. FALSE) and sentence types (**the** vs. **all**).

Figure 11: Example of a filler item from Exp. B1.

4.1 Experiment B1

We tested projection in positive and negative sentences. The materials and procedure were similar to Experiment A1 except that now there was a single group of participants doing a three-valued task (see Fig. 11). We recruited 62 participants, of which 7 were excluded for either having participated in a previous experiment or being non-native speakers of English. This left us with 24 females, 30 males, and one person identifying with neither gender, all aged between 20 and 55.

The results are shown in Fig. 12, and the corresponding statistical results are in Table 6.

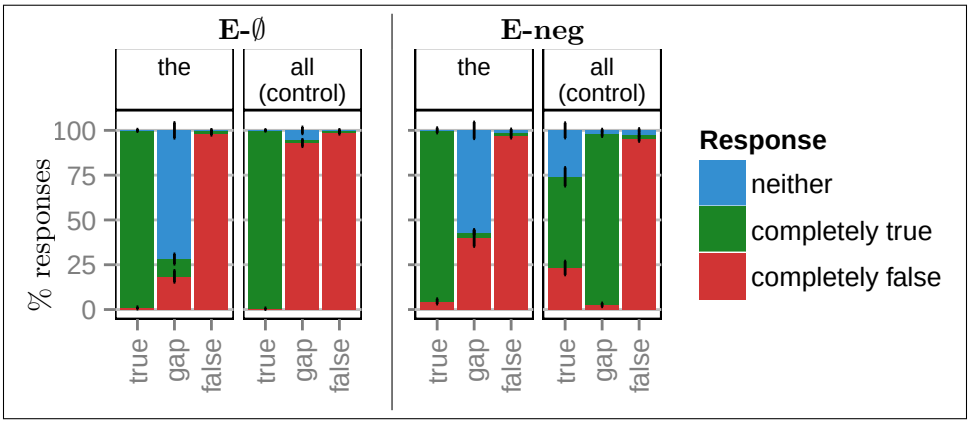


Figure 12: Main results from Exp. B1 in FALSE (0 target-color shape), TRUE (9 target-color shapes) and GAP (1 to 8 target-color shapes) conditions. Average rates of *completely false*-, *completely true*-, and *neither*-responses are stacked on top of each other to add up to 100% of the responses (with error bars at the top of each bar). A high rate of *neither*-responses is evidence of a gap (*neither*-blocks correspond to space left empty in the previous paradigm).

	E-∅			**E-neg**		
	β	χ^2	p	β	χ^2	p
Diag. 1	7.5	72.6	10^{-16}	21	64.1	10^{-15}
Diag. 2	7.2	122	10^{-27}	6.4	96	10^{-22}
Diag. 3	2.9	24	10^{-6}	28	123	10^{-28}

Table 6: Experiment B1: Results of the statistical diagnostics of a gap for positive (**E-∅**) and negative (**E-neg**) sentences. All diagnostics yield significant results for **E-∅** and **E-neg**.

4.2 Experiment B2

We tested projection in simple positive sentences, under *all*, and under *no*. The materials and procedure were similar to Experiment A2, except that now there was a single group of participants doing a three-valued task (see Figure 11). We recruited 61 participants, of which 2 were excluded for being native speakers of a language other than English. This left 36 females and 23 males, aged 18 to 65. The results are given in Fig. 13 and Table 7.

Due to the fact that Exp. B2 used the exact same stimuli as Exp. A2, it contained the same erroneous items for the condition **E-*no*+**FALSE, which were excluded from the analysis. We also find the same inexplicably high rate of *completely false* responses for the controls in **E-*no*+**GAP.

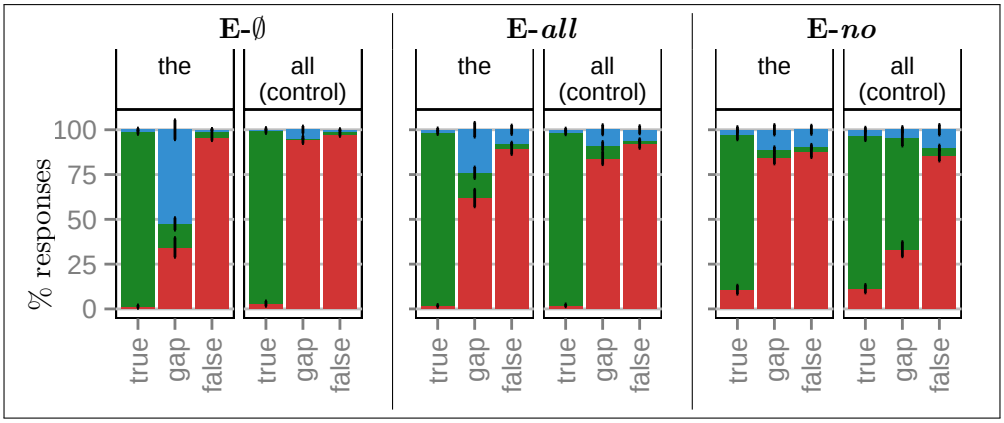


Figure 13: Main results from Exp. B2 (For plotting details see Fig. 12.)

	E-∅			**E-*all***			**E-*no***		
	β	χ^2	p	β	χ^2	p	β	χ^2	p
Diag. 1	5.2	15.2	10^{-4}	3.8	10.8	.0010	−0.68	1.08	.30
Diag. 2	8.5	25.8	10^{-6}	3.5	14.0	.0002	1.4	1.77	.18
Diag. 3	2.7	14.7	10^{-4}	0.8	3.5	.061	0.18	0.31	.58

Table 7: Experiment B2: Results of the statistical diagnostics of a gap for definite descriptions at various levels of embedding: **E-∅**, **E-*all*** and **E-*no***. All diagnostics yield significant results for **E-∅** and **E-*all***, but not for **E-*no***.

4.3 Experiment B3

We tested projection in simple positive sentences, under *all* and under *exactly 3*. The materials and procedure were similar to Experiment A3, except that now there was a single group of participants doing a three-valued task (see Fig. 11). Of a total of 62 participants, 3 were excluded from the analyses for having participated in a previous experiment or being non-native speakers of English. The results, based on 28 female and 31 male subjects aged between 20 and 59, are given in Fig. 14 and Table 8.

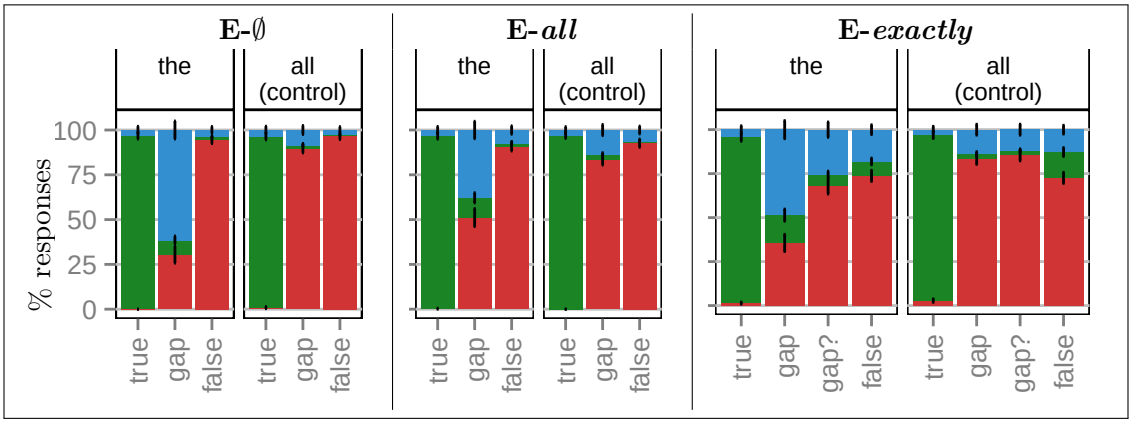


Figure 14: Main results from Exp. B3 (For plotting details see Fig. 12.)

[11] A version of the model with noise (see fn. 4) provides a result which more closely approaches significance: $\beta = .77, \chi^2 = 2.96, p = .085$. This does not change our conviction that there is no strong evidence for a gap in

	E-∅			**E-*all***			**E-*exactly***					
	GAP			GAP			GAP			GAP?		
	β	χ^2	p	β	χ^2	p	β	χ^2	p	β	χ^2	p
Diag. 1	10.5	37	10^{-9}	3.1	22.1	10^{-5}	2.0	7.12	.0076	−0.17	.065	.80
Diag. 2	5.4	28.5	10^{-7}	2.2	15.3	10^{-4}	3.5	20.9	10^{-5}	1.7	5.74	.017
Diag. 3	2.2	11.5	.0007	1.6	7.4	.006	1.8	4.4	.036	0.99	1.8	.18 [11]

Table 8: Experiment B3: Results of the statistical diagnostics of a gap for definite descriptions at various levels of embedding: **E-∅**, **E-*all***, and **E-*exactly***, with two gap conditions GAP and GAP? for the latter. All diagnostics yield significant results except for GAP? with **E-*exactly***.

4.4 Discussion

As shown, the second method we explored in this set of three experiments leads to a striking replication of the previous results, to the point that the graphs look entirely similar: what was empty space above the *completely true* and *completely false* blocks stacked on top of each other in method A, is now filled in by the additional *neither* option in method B. This provides a cross-validation of our way to identify cases which are neither completely true nor completely false.

Before we take stock based on these robust empirical findings, we will extend our ternary methodology a bit further and apply it to slightly different situation-sentence pairs. In doing so we will get rid of a certain problem with **no** sentences (see fn. 10) and test a further type of potential gap situations for non-monotonic environments that will prove to be crucial to the theoretical discussion.

5 Experiments C1-C4: Cleaning up

In this section, we present our final batch of experiments. The purpose of Experiment C2 was to further investigate homogeneity projection from under *no*, fixing the issues that we found with Experiments A2 and B2. After making sure that the new design replicates the previous results for two relevant potential gap conditions with *exactly*, we extend the paradigm to another potential gap-condition with *exactly* in Experiment C4. This will complete the set of data we need to embark on the theoretical discussion in section 6. The experiments were conducted using the ternary method. The methodological innovations vis--vis Experiments B1–3 concern the stimuli, which contained transitive constructions and less abstract (but still schematically depicted) situations, see Sect. 5.2.2 and Figure 15 below.

5.1 Experiment C1

This is an imaginary experiment that replicates experiments A1 and B1, which do not need to be replicated. It is presented in order to maintain the parallelism between the A, B and C series of experiments, but it was never performed.

this condition because: (1) diagnostic 1 does not approach significance (not even in its noisy version); (2) we are not in a situation that involves values at ceiling or floor; (3) the earlier binary method did not produce a significant diagnosis; and (4) the p-value is still high here, especially if correction is being taken into account. In sum, there is no converging evidence for a gap drawing on the three diagnostics and the two methods, contrary to what we find in other conditions.

5.2 Experiment C2

5.2.1 Participants

We recruited 80 participants via Amazon's Mechanical Turk, of whom 6 were excluded for not participating in one of our experiments for the first time or being non-native speakers of English. Of the remaining 74 participants, 31 were female and 43 male, aged between 19 and 67.

5.2.2 Materials

The experimental design deviated from that of Experiment A2 only in minor respects which are all covered by the following description.

Sentences The stimulus sentences contained a proper transitive verb with a definite plural object, with the embedding quantifiers of interest occupying the subject position. We tested two embedding quantifiers, *every* (instead of *all*) and *no*, as shown in (19) and (20); the unembedded variant, which would have had the form in (18), was omitted.

(18) (**E-∅:** Andy found his presents.)

(19) **E-*every*:** Every boy found his presents.

(20) **E-*no*:** No boy found his presents.

For controls, *his* was replaced with *all of his*, and for fillers, we used *few of his* and *most of his*.

Displays The pictures that were used to represent the situation in question were again composed of 4 arrays of 9 objects each. Each cell was associated with one boy and contained 9 pictures of gift boxes. Boxes that were found were colored, while those that were not found were grayed out. This was explained to subjects during in the instructions. A sample item can be seen in Fig. 15.

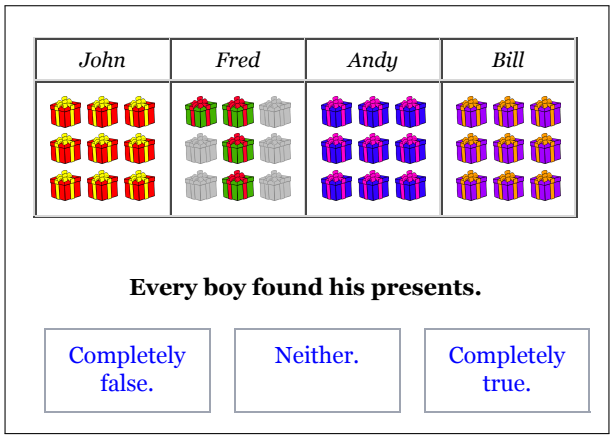


Figure 15: Example of an **E-*every***+GAP item from the C Experiments.

Combinations Every combination of embedding quantifier and truth value was instantiated six times, resulting in 36 target and 36 control items. There were also 24 filler items. The particular configurations followed those from Experiment A2, except that the coding error was corrected and consequently all items for the **E-*no***+GAP condition followed the required pattern.

At the beginning of the experiment, subjects were presented with 12 items that did not count towards the analysis to get them accustomed to the task. These items did include some unembedded cases like (21).

(21) Andy found his presents.

5.2.3 Results

Results are summarised in Fig. 16 and the corresponding statistical analyses are in Table 9. The behavior of **E-*every*** mirrors that of **E-*all*** in Experiment A2, as expected. This time, **E-*no*** does show a gap, which, albeit small, is robust.[12] Furthermore, the controls in the **E-*no*+GAP** condition, which are now grammatical, are no longer evenly divided between *completely true*- and *completely false*-responses. One thing we would expect, but do not find, is a small gap in the controls for the **E-*no*+TRUE** condition reflecting the implicature in (22), which unlike the literal meaning of the sentence is false in this case.

(22) No boy found all of his presents.
⇝ Some boy found some of his presents.

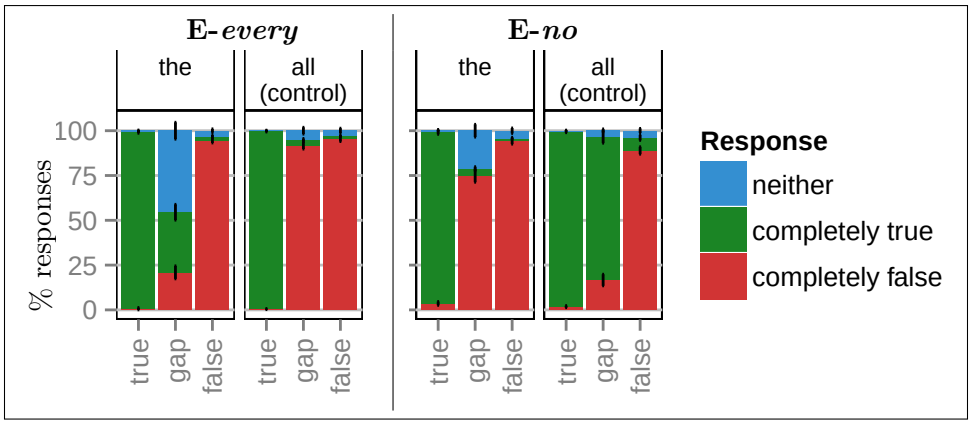


Figure 16: Main results from Exp. C2.

	E-*every*			E-*no*		
	β	χ^2	p	β	χ^2	p
Diag. 1	6.7	26.7	10^{-6}	1.3	8.2	.004
Diag. 2	7.7	35.1	10^{-8}	1.1	5.2	.022
Diag. 3	4.9	4.0	.046	1.6	7.8	.005

Table 9: Experiment C2: Results of the statistical diagnostics of a gap for definite descriptions at various levels of embedding: **E-*every*** and **E-*no***. All diagnostics yield significant results.

5.3 Experiment C3

Experiment C3 was essentially a replication of A3/B3 in the presents-paradigm. Again, it did not include sentences without an embedding quantifier (**E-∅**), but only the two types below.

(23) **E-*all*:** All the boys found their presents.

[12] It also showed up in another, pilot experiment, where the number of boxes varied across cells.

(24) **E-*exactly*:** Exactly 2 of the 4 boys found their presents.

All items were instantiated in analogy to Experiment A3 and B3. Participants recruited via Amazon's Mechanical Turk were 77 in number, of whom 9 were excluded for participating for the second time in one of our experiments or not being a native speaker of English. This left us with 68 participants, 21 women and 47 men, aged between 16 and 69.

The results, in Fig. 17 and Table 10, confirm those of the previous experiments in the verdict that GAP? is not a gap constellation, though one may note that the *at least*-reading, discussed in Sect. 3.4, seems to have disappeared in this experiment.

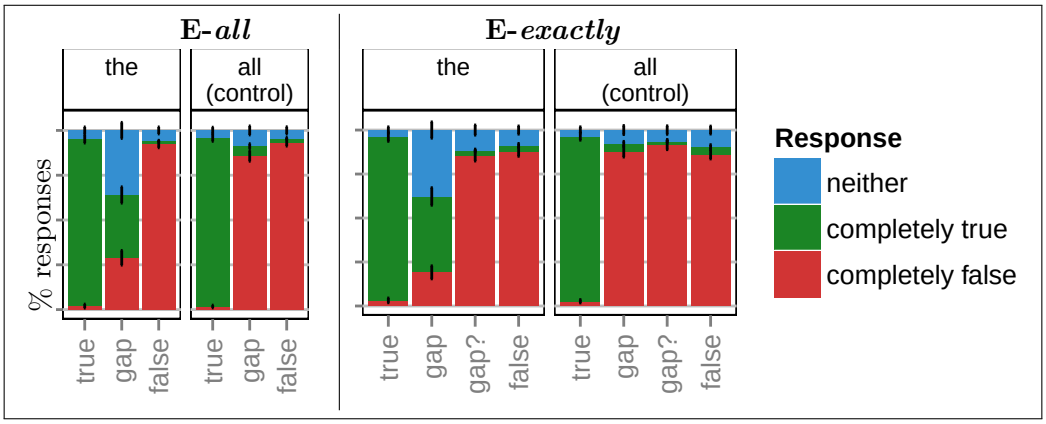


Figure 17: Main results from Exp. C3.

	E-*all*			**E-*exactly***					
	GAP			GAP			GAP?		
	β	χ^2	p	β	χ^2	p	β	χ^2	p
Diag. 1	1.9	18.8	10^{-4}	3.9	21.0	10^{-5}	−0.45	1.4	.23
Diag. 2	1.6	16.4	10^{-4}	7.7	38.8	10^{-9}	.68	2.0	.15
Diag. 3	1.5	9.7	.002	6.6	13.5	.0002	−.01	.02	.88

Table 10: Experiment C3: Results of the statistical diagnostics of a gap for definite descriptions at various levels of embedding: **E-*all*** and **E-*exactly***, with two types of gaps for the latter. All diagnostics yield significant results except for the GAP? condition.

5.4 Experiment C4

This experiment was done exactly like Experiment C3, except that the GAP? conditions were replaced with GAP?? conditions, which had a different configuration: exactly two of the boys found all of their presents, but a third boy found some. We recruited 85 participants via Amazon's Mechanical Turk, of whom one was excluded for not being a native speaker of English. Of the remaining participants, 40 were female, 44 were male, and ages ranged from 19 to 68.

According to our guiding intuition, the new GAP?? condition should result in a gap judgment, and indeed it did, as can be seen in Fig. 18 and Table 11.[13]

[13]Note that the controls with *all* are expected to be judged *false* for GAP, but *true* for GAP?? (as was the case for downward-entailing environments): there are exactly two boys who found all their presents, it's just that a third boy found some of them.

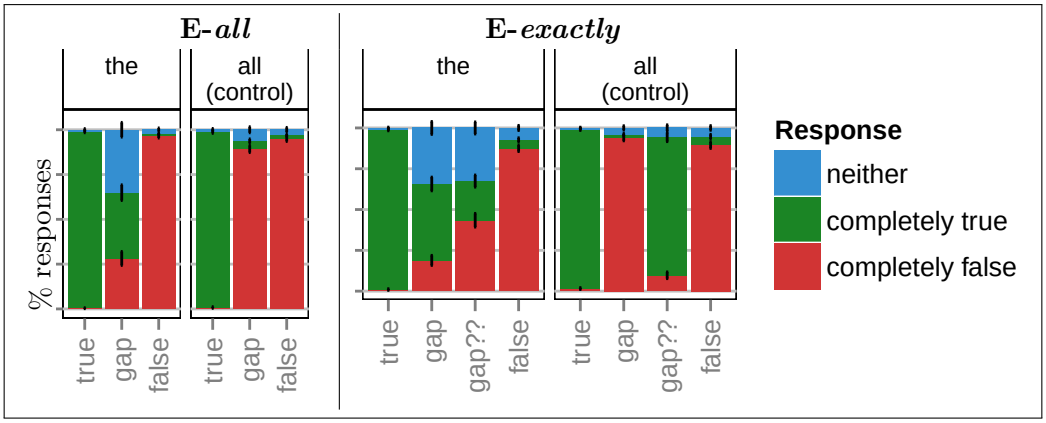


Figure 18: Main results from Exp. C4.

	E-*all*			**E-*exactly***					
	GAP			GAP			GAP??		
	β	χ^2	p	β	χ^2	p	β	χ^2	p
Diag. 1	2.2	17.5	10^{-6}	8.2	16.7	10^{-4}	4.4	13.4	.0002
Diag. 2	1.4	9.9	.002	6.7	36.2	10^{-8}	7.6	49.4	10^{-11}
Diag. 3	1.5	7.6	.006	5.7	6.8	.009	7.0	11.5	.0007

Table 11: Experiment C4: Results of the statistical diagnostics of a gap for definite descriptions at various levels of embedding: **E-*all*** and **E-*exactly***, with two types of gaps for the latter. All diagnostics yield significant results.

6 General discussion: Exploring the landscape of theories

We have described two closely related methods to detect truth-value gaps. These methods generated convergent results regarding the projection pattern for homogeneity, which are summarized in Table 2. Our results provide evidence for a gap in non-embedded environments, under negation, in the scope of *all* and *no*,[14] and in the scope of *exactly* in what we called GAP and GAP?? situations, but not GAP? situations (see Table 12 for a concrete overview of what these different types of gaps are like).

[14]The case of plain negation is uncontroversial, but only Experiment C2 reveals a gap for *no*, while neither A2 nor B2 did. We believe that the results from Experiment C2 are convincing despite the failure to replicate them in A2 and B2, for a number of reasons. First, remember that A2 and B2 used ungrammatical sentences in these conditions (see fn. 10). Second, we suggest that what happened in A2 and B2 is that the intended reading on which *the symbols* would be bound by *none of the cells* was not actually the one that subjects obtained. (We owe thanks to Martin Hackl for a very helpful discussion on this matter.) They rather read *the symbols* as unbound, referring to all symbols in the picture, and *in none of the cells* as quantifying over spatial parts of that symbol plurality. Compare (ia) with (ib):

(i) a. In none of the cells are the symbols red.
b. In none of the corners is the wall red.

It seems that in both sentences, the color adjective is plausibly read existentially: (ib) is false as soon as there is any red spot somewhere around a corner of the wall; it's not necessary that the whole triangle of one corner be red. If this is so, it raises the question why the color adjective was obviously not read existentially in the corresponding sentences with *all*. (ii) was judged neither true nor false in a situation where half of the symbols in every cell were red. Had the color adjective been read existentially in this sentence, it would have been judged true.

(ii) In all the cells, the symbols are red.

We are now in a position to evaluate where these results put us in terms of the current landscape of possible theories. To a large extent, these theories are not yet fully specified and can be spelled out in several variants. We will show how the current set of data puts constraints on these theories.

6.1 Homogeneity as scalar implicature

Magri (2014) presents a theory of homogeneity based on mechanisms used in the domain of scalar implicatures. As a reminder about scalar implicatures, a sentence such as (25a) is supposed to have a stronger alternative (25b). An exhaustivity mechanism (possibly but not necessarily Gricean) may apply and add to the interpretation of (25a) a so-called *scalar implicature* in the form of the negation of this alternative here, resulting in (26).

(25) a. Some of the books are written in Dutch.
b. All of the books are written in Dutch.

(26) **exh**(Some of the books are written in Dutch)
= (25a) and not (25b)
= Some but not all of the books are written in Dutch.

Magri proposes that the same mechanism is active with plain definite descriptions. His theory is based on the following assumptions: (i) (27) has an existential literal meaning, just like (25a); (ii) the alternatives to the definite include *some*, but not *all*; and (iii) the alternatives to *some* include *all*, but not the definite. The audacious proposal that follows from this is that (27) has an existential literal meaning, that no enrichment is obtained if **exh** is applied once (because the two alternatives (25a) and (27) are equivalent), but that (27) acquires a universal meaning upon a second application of exhaustification (as the result of the negation of the exhaustified (26)). The computation is given in (29).

(27) The books are written in Dutch.

(28) **exh**(The books are written in Dutch)
= The books are written in Dutch.

(29) **exh**(**exh**(The books are written in Dutch))
= **exh**(The books are written in Dutch) and not **exh**(some ...)
= Some of the books ... and not (some but not all ...)
= All of the books are written in Dutch.

In general, the exhaustivity operator is taken to negate non-weaker (as opposed to stronger) alternatives, as has become largely standard by now. As a result, for a sentence φ(the) containing a definite description, exhaustification will strengthen the overall meaning unless the environment φ is downward-entailing, as follows:

We do not know why, but it seems to us that empirically, walls with red corners again behave analogously: (iiia) is only true if the whole triangle that makes up one corner is red; and (iiib) requires the same of all corners.

(iii) a. In the corner, the wall is red.
b. In all the corners, the wall is read.

We do not have any further explanation for this curious pattern of interaction between topicalisation of the quantified prepositional phrase, negative inversion, and existential vs. homogeneously universal readings of color adjectives. Overall, various complications cast doubt on any strong conclusion from the failure of experiments A2 and B2 to find a gap with *no*. What we have to leave unexplained is that the gap with *no* is somewhat smaller than in other conditions even in Experiment C2.

(30) **exh**(**exh**(φ(the))) =

a. φ(some) *if $\varphi(\cdot)$ is downward-entailing*
b. φ(some) and φ(all) *otherwise*

6.1.1 Processing data

It is useful to consider the import of processing data at this point. At first, they may seem to present a problem for the conceptual underpinnings of Magri's theory. Magri argues that maximal readings arise as scalar implicatures. But in similar tasks, scalar implicatures show slow response times (see Bott and Noveck 2004 and others), while maximal readings show fast response times (Schwarz 2013a). However, other processing studies on phenomena that are allegedly related to scalar implicatures have also found the reverse processing results. Work by Chemla and Bott (2014) offers a striking point of comparison: they found similar results for free choice inferences (fast processing, in short), even though these inferences are alleged cases of implicatures based on double exhaustification, just like homogeneity would be in Magri's theory.

6.1.2 Gaps in an implicature-based approach

Without further assumptions, Magri's theory does not predict any gap in our sense, as it is only concerned with the truth conditions of sentences. However, in another work, Magri (2009) defends the idea that if a sentence is literally true but false when enriched with its scalar implicatures, then a feeling of 'oddness' arises. This view leads Magri to an account of the deviance of sentences such as *Some Italians live in a warm country.*

More generally, one can imagine several ways in which an imperfect truth-value judgment may arise, as a conflict between the truth values of either different parts of the meaning of a sentence (semantic content, communicated content) or different readings of it. Let us make some of these scenarios concrete, explain their possible motivations and briefly give a first description of the empirical predictions they could lead to. We note that the operative notion of gappiness is quite broad here, in that it predicts that the gap judgments we obtained should be replicable with run-of-the-mill kinds of scalar implicatures, for instance.[15] In fact, we may see evidence of scalar implicatures generating gaps in our studies. Consider the sentence *Not all triangles are green.* It may have a scalar implicature that *some* triangles are green. In experiments A1 and B1, we tested exactly such sentences as control **all**-sentences and the TRUE condition was such that the sentence was literally true but its scalar implicature was not. We observe some evidence of a gap there. For instance, looking at the second panel of Fig. 12, we see a 25% share of *neither* responses in this condition. As a challenge to the scalar implicature theory, however, note that the size of this scalar implicature generated gap seems smaller than the size of other homogeneity gaps we find and there is a large proportion of *completely true*-responses based on the literal meaning (whereas in homogeneity-based gaps, *completely false*-responses by far outnumber *completely true*-responses).

In determining where exactly a scalar implicature approach might predict gaps, it is useful to first have a look at the various building blocks of meaning that could be used here. Given a sentence φ(the) containing a definite description, where $\varphi(\cdot)$ is not downward-entailing, those blocks of meaning are as follows:[16]

[15] Interestingly, these are the kinds of examples used in the processing literature on scalar implicatures, as in *Some elephants are mammals.*

[16] Note that the exhaustification view does not derive implicatures in the form of propositions to be added to the literal/non-exhaustified meaning. Instead we simply obtain a composite meaning, putting together all the pieces without structure. If **exh** is applied globally, one could define implicatures as the negations of all

(31) a. Literal meaning: φ(some)
b. Implicature: $\neg(\varphi(\text{some}) \wedge \neg\varphi(\text{all})) \equiv \neg\varphi(\text{some}) \vee \varphi(\text{all})$
c. Globally exhaustified meaning:
$\mathbf{exh}(\mathbf{exh}(\varphi(\text{the}))) = \varphi(\text{some})$ and $\varphi(\text{all})$
d. Locally exhaustified meaning:
$\varphi(\mathbf{exh}(\mathbf{exh}(\text{the}))) = \varphi(\text{all})$

Using these elements, one could imagine a number of mismatches that may correspond to a gap, hence creating several variants of a scalar implicature theory of homogeneity:

(SI_1) Gap = literal meaning true and implicature false.

This option aligns with Magri's (2009) condition for scalar items to generate oddness.

(SI_2) Gap = literal meaning and implicature have conflicting truth values.

This option predicts additional gap cases, compared with (SI_1), for non-monotonic contexts, where the implicature and the literal meaning may be logically independent (see below).

(SI_3) Gap = two parses with conflicting truth values; only global exhaustification is available.

If the insertion of the exhaustification operator is optional, we are facing cases of ambiguity which could create a gap-like judgment (*neither completely false nor completely true*) if the two readings push in different directions. This option is empirically equivalent to (SI_1), since the globally exhaustified meaning is the conjunction of the literal meaning and the implicature. It may arise from a different principle, though, which would predict that our methodology should turn up gap-judgments when applied to ambiguous sentences in general.

(SI_4) Gap = two parses with conflicting truth values; local exhaustification is (also) available.

The motivation behind this option is the same as for the previous one, assuming that arguments for the existence of local exhaustification are effective. This option is empirically equivalent to (SI_2) except for downward-entailing contexts (if we accept that local exhaustification is possible in downward-entailing contexts).[17]

In the following, we will examine our empirical findings for various embedding contexts with respect to how they would guide the choice between these options.

6.1.3 Scalar implicature approach: assessment based on the empirical findings

Embedding under *all* For the case of embedding in the scope of *all*, all of the options above align, since the implicature, the globally exhaustivised meaning, and the locally exhaustivised meaning all amount to the same thing. Thus, for a sentence like (32) a gap is predicted in a situation where all teachers like at least some of their students, but not all teachers like all of their students.

(32) All the teachers like their students.

the alternatives that end up being negated, or as the weakest propositions that need to be added to the non-exhaustified meaning to obtain the exhaustified meaning, but in fact these definitions won't go far for many cases of local exhaustification.

[17] Taking into account parses with one global and one local **exh** would not add any new prediction.

This is exactly the way we determined the GAP condition in our experiments, and the theoretical predictions were borne out: a gap was found in these conditions, but the sentences with *all* in lieu of the definite were judged false as soon as one cell contained any non-target-color symbols at all.

Downward-entailing environments Let us first look at the variants of the scalar implicature theory (SI_1), (SI_2), and (SI_3). A sentence with a definite under negation or in the scope of a downward-entailing quantifier such as *no*, as in (33) and (34), should have the following properties: first, the definite effectively has an existential interpretation, and second, there is no implicature. The reason for this is that in a downward-entailing environment, an existential leads to a stronger meaning for the whole sentence than a universal. Thus, exhaustification of *some* is vacuous here because there is no non-weaker alternative to exclude as an implicature; and since *some* has no implicature, the plain definite does not have an implicature, either, and we are just left with its literal (existential) meaning.

(33) The books aren't written in Dutch.

(34) No teacher likes his students.

Leaving gaps aside for a moment, this outcome is satisfying: (33) does indeed mean that no book was written in Dutch (see initial discussion of (2b)), and similarly, (34) is completely true only if no teacher likes any of his or her students. But the absence of an implicature also predicts that there is no gap, because a situation where the literal meaning is true and the implicature is false cannot arise. This is problematic because we find clear evidence of a gap for plain negation (see Experiments A1 and B1) and a small but robust gap for *no* in Experiment C2.

For plain negation, a way out of this problem would be to assume that (33) has a parse where the definite takes scope over negation, i. e. 'The books are [not written in Dutch]'.[18] Then the definite is in an upward-entailing environment, and (33) has the literal meaning in (35a) and the implicature in (35b). On this parse, the sentence has a gap that is the same as that for its positive counterpart.

(35) a. Some of the books are [not written in Dutch].
b. All of the books are [not written in Dutch].

For a sentence with *no* like (34), however, this solution does not help to predict a gap, since the definite contains a variable bound by the negative quantifier, and it is furthermore known that definites do not take distributive scope over quantifiers (Steedman 2012), and the gap found for *no* in Experiment C2 thus remains unexplained.

Let us now look at option (SI_4), which allows local exhaustification under negation. A gap is predicted for both sentential negation and *no*, since the definite description is then locally in an upward-entailing contexts. The sentence in (36) is predicted to display a gap if (36a) and (36b) have different truth values, i. e. if some, but not all ,of the books are in Dutch.

(36) The books aren't in Dutch.
a. not [the books are in Dutch] = no book is in Dutch
b. not **exh**(the books are in Dutch) = not all the books are in Dutch

[18] Note that this does not require distributivity in the derived position; distributivity can still happen below negation. Instead of denoting a plurality, the definite plural then just quantifies over pluralities without being inherently distributive, which would allow it to take derived scope without standing in contradiction to Steedman (2012).

Analogously, (37) should have a gap when no teacher likes all of his students, but some teacher likes some of his students.

(37) No teacher likes his students.

a. [no teacher] [likes his students]
= no teacher likes any of his students

b. [no teacher] **exh**(likes his students)
= no teacher likes all of his students

In sum, version (SI_4) of the implicature theory, the one which accepts local exhaustification, seems to be what is needed if one wants to cover all the facts so far. But this is not an appealing option, since there is independent evidence that local exhaustification is not available in downward-entailing contexts (Chierchia, Fox, and Spector 2012).

Non-monotonic environments Finally, we need to take a look at non-monotonic environments as created in the scope of *exactly*. Magri's exhaustivity operator includes the negation not only of stronger alternatives, but also of logically independent alternatives. As a result, unlike what happened in downward-entailing environments, there is an effect for *some* embedded under *exactly*. For a sentence such as (38), we obtain the result computed in (39).

(38) Exactly 2 teachers like their students.

(39) **exh**(**exh**(Exactly 2 teachers like their students))
= **exh**(Ex2 . . . the)
and not [**exh**(Ex2 . . . some)]
= Ex2 . . . some
and not [Ex2 . . . some and not(Ex2 . . . all)]
= Ex2 . . . some and Ex2 . . . all.
= Exactly 2 teachers like all of their students,
and no other teacher likes any of his.

It is useful to look at a range of configurations and the predictions that the different versions of the theory make for them. Assume that we have three teachers, Bill, Mary, and Sue. We can now have a look at various situations. Table 12 shows, for each of several situations, how many of their students each teacher likes in that situation.

				Exhaustification			(SI_1)/(SI_3)	(SI_2)/(SI_4)	Corresponding	Gap
	Bill	*Mary*	*Sue*	lit	loc	glob	lit≠glob	lit≠loc	Conditions	found?
s_1	all	all	none	T	T	T	T	T	TRUE	no
s_2	all	all	all	F	F	F	F	F	FALSE	no
s_3	all	none	none	F	F	F	F	F	FALSE	no
s_4	all	half	none	T	F	F	gap	gap	GAP	yes
s_5	all	half	half	F	F	F	F	F	GAP?	no
s_6	all	all	half	F	T	F	F	gap	GAP??	yes

Table 12: Various situations with corresponding truth values of the different parses — lit(eral, i.e. no exhaustification), loc(al exhaustification), glob(al exhaustification) — and predictions of the different construals of a gap in an implicature approach, and corresponding conditions and results from our experiments.

Sentence (38) is obviously true in s_1, and obviously false in s_2 and s_3. Situation s_4 is predicted to be a gap under any notion of 'gap' in the theory. It is the case that exactly 2

teachers like some of their students, but the implicature that exactly 2 teachers love all of their students is false: only John loves all of his students. In s_5, both the literal meaning and the implicature are false, and so the sentence is again predicted to be simply false. Finally, in s_6, we have the reverse situation: the literal meaning is false, but the implicature, which is logically independent from it, is true. Here, the prediction depends on what exactly constitutes a gap: truth of the literal meaning and falsity of the implicature, or, more broadly, any difference in truth value between the literal meaning and the implicature. In the first case, the sentence is predicted to be simply false in s_6; in the second, it is predicted to be gappy.

In our experiments, the TRUE conditions correspond to s_1 and the FALSE conditions to s_2 or s_3. The GAP condition corresponds to the uncontroversial s_4, while the GAP? condition tests s_5. On all the tested cases, the predictions of the theory are thus borne out, as a gap was found in GAP, but not in GAP?. Finally, s_6 corresponds to the GAP?? condition as tested in Experiment C4, where it was found to be a genuine gap. This constrains the ways in which the account for gaps on an implicature theory would have to be spelled out. In particular, unless local exhaustification is used, the implicature itself has to be singled out as a meaning component and compared with the non-exhaustified meaning. If they differ in truth value, then a gap judgment arises.

6.2 Homogeneity as vagueness: The supervaluations approach

Spector (2013) proposes a different approach to homogeneity, which is further developed in Križ and Spector (2015). The underlying idea is that a definite is, in some sense, vague or underspecified as to its quantificational force (which may be either existential or universal), and that a sentence is true if it is true no matter how we resolve this, and false if it is false no matter how the definite is interpreted. Thus the system is related in spirit to supervaluationist theories of vagueness.

To illustrate, assume that the two possible interpretations for the definite are just an indefinite and a universal. Then the simple sentence (40) is true if both (40a) and (40b) are true — that is, if all children are alive — and false if both (40a) and (40b) are false — that is, if all children are dead.

(40) The children are alive.
   a. Some children are alive.
   b. All children are alive.

Although the conceptual basis is completely different, it is convenient to note here that this system is empirically equivalent to variant (SI_4) of the scalar implicature theory (see Sect. 6.1.2) as far as projection is concerned: using the terminology from the scalar implicature approach, the supervaluationist theory identifies gaps as cases in which the sentence does not have the same truth value as its locally exhaustified version, since the locally exhaustified version is just the same sentence where *the* has been replaced with *all*. Consequently, this theory predicts the right distribution of gaps over all the conditions we tested, just like (SI_4) does, but without the conceptual weaknesses of (SI_4).

Just as with scalar implicature theories, there are variants of the supervaluations approach. One possibility is to introduce more subtle distinctions between candidate meanings (i.e. beyond simple existential and universal meanings), and some of these predict additional gap judgments with non-monotonic quantifiers, namely in the GAP? condition.[19] Given that we ended up not finding a gap in that condition (cf. Experiments A2, B2, C2), we refer to Križ and Spector (2015)

[19] As an aside, note that many quantified trivalent logics such as Łukasiewicz logic or Gödel logic, or a natural extension of Strong Kleene logic to a first order language, make the same prediction.

for a more complete discussion of this theory, whose motivation comes from non-maximality effects of the kind presented in Sect. 1.3, rather than from projection facts.

6.3 Homogeneity as a presupposition

The historically oldest explicit position on homogeneity is that it is a presupposition (Schwarzschild 1994, Löbner 2000, Gajewski 2005). These authors assumed that definite plurals have universal force. A variant of the presupposition theory in which the definite has existential force may have the advantage of making non-maximality easier to understand and would provide slightly weaker predictions for our target cases; but for the sake of simplicity, we will not explore this ramification further. The main assumption then is that predication of a (distributive) predicate P of a plurality x presupposes that either all parts of x are P or no part of x is P. So (41a) and (42b) are taken to presuppose that the set of triangles is green-homogeneous: either all triangles are green or none are green (and these sentences should assert opposite claims about whether there are in fact green triangles).

(41) a. The triangles are green.
b. The triangles are not green.

A presuppositional gap is thus predicted for positive and negative sentences such as (41a) and (41b), and this is coherent with the results of Experiments A1 and B1.

The actual patterns of presupposition projection from under quantifiers are not uncontroversial (see, e.g., Chemla 2009 for discussion and relevant data). Thus, we cannot make a decisive argument for or against the theory that homogeneity is a presupposition. But we may note that if it is a presupposition, our results put relevant constraints on the theory of presupposition projection; and the soundness of these constraints can then be evaluated.

First, we observe that our results exclude a theory on which presuppositions project universally from under *all*/*every* and *no*. Sentence (42) certainly *entails* that the symbols in all cells are green-homogeneous, but it cannot presuppose this, since we found that it is false as soon as there is one cell in which no symbol is green — the other cells need not be green-homogeneous in this case.

(42) In all the cells, the symbols are green.

Second, predicting a presupposition failure in the GAP condition, but not in the GAP? condition, seems to be a challenge for some theories of presupposition projection. In short, one would need a theory that predicts a presupposition that is neither universal (to avoid predicting a gap in GAP?) nor too weak (to actually predict a gap in GAP and GAP??). George's (2008a; 2008b; 2008c) family of presupposition projection theories, based on trivalent logic, includes some particular versions that reach this goal and make the exact same predictions as the supervaluations approach discussed above.

In sum, it is, at present, not possible to reach a conclusive verdict about the status of homogeneity as a presupposition solely on the basis of the homogeneity projection data we presented and our current understanding of presupposition projection. Two avenues would be worth pursuing in this matter. First, the application of our own methodology to presuppositions would have the intrinsic interest to detect presupposition gaps as such. Second, Zehr (2014a) is developing tests that presumably distinguish between presuppositions and borderline cases of vague predicates, which could be applied to homogeneity phenomena to assess whether they pattern with presuppositions or vagueness.

6.4 Summary: Our findings and current theories

Let us summarize our data and how they bear on existing theories of homogeneity. We found a gap for plainly positive and negative sentences. We found a gap for more complex upward-entailing environments (scope of *all*) and downward entailing environments (scope of *no*). We also found a gap for non-monotonic environments too (scope of *exactly n*): in the GAP condition (where exactly n cells contained target-objects, although not all of them were *homogeneously* filled up with target objects), in the GAP?? condition (where exactly n cells contained *only* target target objects, while another cell contained some), but not in the GAP? condition (where more than n cells contained target objects, while less than n cells contained such objects exclusively).

Three major conceptual approaches to homogeneity have been proposed.Each of these approaches comes in several variants that we have tried to spell out. To meet the current empirical findings, the scalar implicature approach requires the possibility of local exhaustification in downward-entailing contexts, the supervaluation approach requires a restricted set of candidate meanings (which may be insufficient for other purposes, see Križ & Spector 2015) and the presupposition approach requires a particular projection theory, probably one like George's (2008b) theory.

7 Conclusion

In this paper, we have developed two experimental methods for diagnosing conditions under which a sentence is neither (clearly) true nor (clearly) false. In method A, a first group of participants judge whether a sentence is *completely true* or not and a second group judges whether it is *completely false* or not. Comparing the results, we assess the cases in which the sentence is neither *completely true* nor *completely false*. In method B, a single group of participants judge whether a sentence is *completely true*, *completely false* or *neither*. We demonstrated that small details of experimental design are important for obtaining a reliable method (e.g., use of the adverb, form of the third option, mouse clicks on the response choices rather than pressing keys). We hope that these efforts can provide a strong methodological basis for experimental studies of phenomena involving truth-value gaps: presuppositions, vagueness, and homogeneity, and we have argued that a joint investigation of these phenomena (as undertaken in Zehr 2014a and Spector 2015 for presupposition and vagueness) would also be welcome.

We applied our methods to homogeneity of sentences with definite plurals, in particular to the question of how the homogeneity of embedded definite plurals projects from the scope of quantifiers, and strove to present as complete a picture of this as possible. A close investigation turned up that current theories of homogeneity still have free parameters to be specified, and that our results constrain them in a particular way, as well as setting a bar to clear for future theoretical developments.

Appendix: Item details

Table 13 details the target and control items used in the experiments. Table 14 contains the filler items. These tables are valid for all experiments that used the relevant conditions (i.e. identical conditions were always instantiated identically across experiments, unless otherwise noted).

	E-∅	**E-*all***			**E-*no***			**E-*exactly***		
FALSE	0	0990	9770	9079	9009	5009	0920	0900	9990	0000
		9440	4044	4990	(0550)	(5055)	(0220)	4000	9999	9994
					[5009]	[5009]	[9909]			
GAP	8 6	9929	9939	9499	0070	0060	0030	9200	0490	7090
	4 2	9969	9799	9995	0060	0500	0200	3300	0550	0808
TRUE	9		9999			0000		9900	9090	9009
								0909	0099	0990
GAP?	–		–			–		9202	3390	0494
								5599	9660	7907
GAP??	–		–			–		9209	3990	0994
								9590	9960	7909

Table 13: Target (**the**) and control (**all**) items. A number string *abcd* is to be read as indicating that in cell 1, *a* of the nine symbols were of the target color mentioned in the sentence, in cell 2 it was *b* symbols, etc. Items in round brackets were faulty items from A2 and B2 (excluded from the analysis), which were replaced by the items in square brackets in C2.

	E-∅	**E-*all***	**E-*no***	**E-*exactly***
uncontroversially true for *most*	7	7777	0220	7007
and false for *few*			0330	
uncontroversially true for *few*	2	2222	7007	0220
and false for *most*		3333		0330
undecided (true, unanalysed fillers)	5	0000	0000	0000
	9	5555	0505	0505
		9999	9900	9900

Table 14: Filler items with **few** and **most**. Uncontroversially true and false items were used to filter participants for accuracy above 75%.

Acknowledgments

We would like to thank Alexandre Cremers, Paul Égré, Martin Hackl, Irene Heim, Jeremy Kuhn, Florian Schwarz, Benjamin Spector and Jérémy Zehr for critical discussion and guidance. We would also like to thank the team of editor, reviewer and copy editor (Christine Bartels) at Natural Language Semantics. This research has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n.313610 and was supported by ANR-10-IDEX-0001-02 PSL* and ANR-10-LABX-0087 IEC.

References

Abrusán, M., & Szendrői, K. (2013). Experimenting with the king of France. *Semantics & Pragmatics*, *6*(10), 1–43. doi: 10.3765/sp.6.10

Alxatib, S., & Pelletier, J. (2011). On the psychology of truth-gaps. In R. Nouwen, R. van Rooij, & H.-C. Schmitz (Eds.), *Vagueness and communication* (Vol. 6517, pp. 13–36). Heidelberg: Springer.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255-278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4.* ArXiV e-pring. Retrieved from `http://arxiv.org/abs/1406.5823`

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, *51*(3), 437–457.

Breheny, R. (2005). Exhaustivity, homogeneity, and definiteness. In P. Dekker & M. Franke (Eds.), *Proceedings of the fifth Amsterdam Colloquium* (pp. 59–65).

Brisson, C. (1998). *Distributivity, Maximality, and Floating Quantifiers.* Unpublished doctoral dissertation, Rutgers.

Büring, D., & Križ, M. (2013). It's that and that's it! Exhaustivity and homogeneity presuppositions in clefts (and definites). *Semantics & Pragmatics*, *6*, 1–29.

Burnett, H. (2013, November). *Vague predication, definiteness, and distributivity.* Talk at the Substructural Approaches to Paradox workshop, University of Barcelona.

Chemla, E. (2009). Presuppositions of quantified sentences: Experimental data. *Natural Language Semantics*, *17*(4), 299–340.

Chemla, E., & Bott, L. (2014). Processing inferences at the semantics/pragmatics frontier: disjunctions and *free choice*. *Cognition*, *130*(3), 380-396. doi: 10.1016/j.cognition.2013.11.013

Chemla, E., & George, B. (2015). Can we agree about 'agree'? *The Review of Philosophy and Psychology*. (Accepted with minor revisions)

Chierchia, G., Fox, D., & Spector, B. (2012). Scalar implicatures as a grammatical phenomenon. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics: An international handbook of natural language meaning* (Vol. 3, pp. 2297–2331). Berlin: Mouton de Gruyter.

Egré, P., de Gardelle, V., & Ripley, D. (2013). Vagueness and order effects in color categorization. *Journal of Logic, Language and Information*, *22*(4), 391–420.

Fodor, J. D. (1970). *The linguistic description of opaque contexts.* Unpublished doctoral dissertation, MIT, Cambridge, Mass.

Frege, G. (1892). *On sense and reference.*

Gajewski, J. (2005). *Neg-raising: Polarity and Presupposition.* Unpublished doctoral dissertation, MIT.

George, B. (2008a). A new predictive theory of presupposition projection. In T. Friedman & S. Ito (Eds.), *Proceedings of SALT 18* (pp. 358–375). Ithaca, NY: CLC Publications.

George, B. (2008b). *Prediction presupposition projection: Some alternatives in the strong Kleene tradition.* Manuscript, UCLA.

George, B. (2008c). *Presupposition repairs: a static, trivalent approach to predicting projection.* Unpublished master's thesis, UCLA.

Heim, I. (1983). On the projection problem for presuppositions. In *Proceedings of WCCFL 2* (Vols. Barlow, Michael and Flickinger, Daniel and Wescoat, Michael, pp. 114–125). Stanford University.

Karttunen, L., & Peters, S. (1979). Conventional implicature. In C.-K. Oh & D. Dinneen (Eds.), *Syntax and Semantics 11: Presupposition.* New York: Academic Press.

Krifka, M. (1996). Pragmatic strengthening in donkey sentences and plural predications. In T. Galloway & J. Spence (Eds.), *Proceedings of SALT 6* (pp. 136–153). Ithaca, NY: CLC Publications.

Križ, M., & Spector, B. (2015). *A theory of homogeneous plural predication.* Manuscript, University of Vienna and Institut Jean Nicod.

Lasersohn, P. (1999). Pragmatic halos. *Language*, *3*(75), 522-551.

Löbner, S. (1987). The Conceptual Nature of Natural Language Quantification. In I. Rusza & A. Szabolcsi (Eds.), *Proceedings of the '87 Debrecen Symposion on Logic and Language* (pp. 81–94). Budapest: Akadémiai Kiadó.

Löbner, S. (2000). Polarity in natural language: predication, quantification and negation in particular and characterizing sentences. *Linguistics and Philosophy*, *23*, 213-308.

Magri, G. (2009). A theory of individual-level predicates based on blind mandatory scalar implicatures. *Natural Language Semantics*, *17*(3), 245–297.

Magri, G. (2014). An account for the homogeneity effects triggered by plural definites and conjunction based on double strengthening. In S. Pistoia Reda (Ed.), *Semantics, pragmatics and the case of scalar implicatures* (pp. 99–145). Basingstoke: Palgrave Macmillan.

Malamud, S. (2012). The meaning of plural definites: A decision-theoretic approach. *Semantics & Pragmatics*, *5*, 1–58.

Marty, P., Chemla, E., & Spector, B. (2014). *Phantom readings: the case of modified numerals.* (Manuscript, IJN, LSCP, MIT)

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Ripley, D. (2011). Contradictions at the borders. In R. Nouwen, R. van Rooij, & H.-C. Schmitz (Eds.), *Vagueness and communication* (Vol. 6517, pp. 169–188). Heidelberg: Springer.

Sauerland, U. (2011). Vagueness in language: The case againast fuzzy logic revisited. In P. Cintula, C. Fermüller, L. Godo, & P. Hájek (Eds.), *Understanding vagueness - logical, philosophical and linguistic perspectives* (pp. 185–198). London: College Publications.

Schwarz, F. (2013a). Maximality and definite plurals: Experimental evidence. In E. Chemla, V. Homer, & G. Winterstein (Eds.), *Proceedings of Sinn und Bedeutung 17* (pp. 509–526).

Schwarz, F. (2013b). *False but slow: Rejecting statements with non-referring definites.* Manuscript, University of Pennsylvania.

Schwarzschild, R. (1994). Plurals, presuppositions and the sources of distributivity. *Natural Language Semantics*, *2*(3), 201–248.

Serchuk, P., Hargreaves, I., & Zach, R. (2011). Vagueness, logic and use: Four experimental studies on vagueness. *Mind & Language*, *26*(5), 540–573.

Spector, B. (2013, September). *Homogeneity and plurals: From the strongest meaning hypothesis to supervaluations.* Presented at Sinn und Bedeutung 18, University of the Basque Country.

Spector, B. (2015). Multivalent semantics for vagueness and presupposition. *Topoi*.

Steedman, M. (2012). *Taking scope*. Cambridge, Mass.: MIT Press.

Tarski, A. (1935). Zur grundlegung der boole'schen algebra I. *Fundamenta Mathematicae*, *24*(1), 177–198.

Tarski, A. (1956). The concept of truth in formalized languages. *Logic, semantics, metamathematics*, *2*, 152–278.

Wickham, H. (2007). Reshaping data with the `reshape` package. *Journal of Statistical Software*, *21*(12), 1–20. Retrieved from `http://www.jstatsoft.org/v21/i12/`

Wickham, H. (2009). *`ggplot2`: elegant graphics for data analysis.* New York: Springer. Retrieved from `http://had.co.nz/ggplot2/book`

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*(1), 1–29. Retrieved from `http://www.jstatsoft.org/v40/i01/`

Zehr, J. (2014a). *Truth-judgments for vagueness and presupposition.* (Presented at the Experimental Philosophy group at Institut Jean Nicod, Paris)

Zehr, J. (2014b). *Vagueness, presupposition and truth-value judgments.* Unpublished doctoral dissertation, École Normale Supérieure de Paris.