Not so squeamish, Your Majesty: The presuppositions of singular definites between existence and uniqueness*

Maik Thalmann
University of Göttingen

July 17, 2025

This study examines the semantic behavior of singular definite descriptions, focusing on the inferences of existence and uniqueness. Using trivalent truth-value judgment tasks, we found that uniqueness violations pattern with homogeneity violations—distinct from true or false controls—while existence violations align with false controls. These results challenge both Russellian and Fregean accounts, which predict that both inferences should behave the same, as mere entailments or presuppositions, respectively. Instead, our findings are broadly compatible with Coppock & Beaver (2015). Lastly, we introduce the *variance hypothesis*, proposing that presupposition failure leads to increased response variability—interpreted as a marker of uncertainty or squeamishness (Strawson 1964). This was borne out for uniqueness and homogeneity violations, suggesting that variance, often dismissed as mere noise, can serve as a useful diagnostic in identifying presupposition failure.

Keywords: definiteness, uniqueness, existence, presupposition failure, experimental semantics

1 Introduction

Squeamishness, a term due to Strawson (1964), describes the sense of momentary discomfiture speakers experience when encountering presupposition failure—cases where the admittance conditions of an utterance are not supported by the context. Presuppositions, which are traditionally seen as requirements for discourse felicity, when violated, cause speakers unsure how to proceed. And since meeting presuppositional requirements is a prerequisite for deciding between truth and falsity, the ensuing squeamishness may be a diagnostic for truth-value gaps.

The main focus of this paper will be on the presuppositions triggered by singular definites. These are commonly assumed to come with two kinds of presuppositions (for a prominent example, consider Heim & Kratzer 1998 a.m.o.): existence and uniqueness. The former is the presupposition that there is an individual that satisfies the description, while the latter is the presupposition that there be at most one such individual. When these presuppositions are not jointly satisfied—when there is not exactly one predicate-satisfying individual—, the context falls into the truth-value gap and elicits squeamishness, or

*For feedback and guidance at various stages of the project, I want to thank Andrea Matticchio, Clemens Mayr, and Thomas Weskott, as well as the members of the reading group for empirical lingustics (LEGEBLING) in Göttingen, the audiences at the Oberseminar English Linguistics in Göttingen, the collaborative workshop with the RTG on nominal modification Frankfurt, and at CONSOLE 32, London. For her feedback, checking both the linguistic and the visual materials in the experiments, and help with carrying out the experiments, I owe Evelyn Ovsjannikov a thank you and probably several apologies. The responsibility for the remaining errors lies solely with me. This project was carried out within the RTG 2636 "Form-Meaning Mismatches," funded by the *Deutsche Forschungsgemeinschaft*, project number: 429844083.

an intuition of undefinedness, in discourse participants. As an illustrative example of this feeling, consider the well-worn matter of the missing majesty from Russell (1905: p. 483):[1]

(1) The king of France is bald. (#)

Despite the fame of the case in (1) and the relative consensus regarding the presuppositional requirements triggered by singular definites like *the king of France*, the presupposition status of existence and uniqueness remains not without its detractors and with scarce amounts of experimental backing. Here, I will try to get closer to answering the question of what exactly lies in the truth-value gap of singular definites and how such gaps behave in speakers' judgments when probed through experimental rather than introspective means.

In what follows, I examine the interpretive profile of singular definites by reviewing three views of their presuppositional behavior: (i) the Russellian approach (cf. Russell 1905), (ii) the Fregean view (cf. Frege 1997), and (iii) the much more recent Weak Fregean account proposed by Coppock & Beaver (2015). These will inform the core hypotheses our later experiments are designed to test.

I also explore the secondary question of what happens when presuppositions fail. While classical theory predicts undefinedness across the board, some failures (especially of existence) appear to be interpreted as outright false sentences, contrary traditional commitments regarding truth-value gaps. A pertinent example of such non-catastrophic presupposition failure from Strawson (1964: p. 114) is shown in (2). Two leading proposals to account for these asymmetries involve explanations driven by topicality (Strawson 1964) and verification strategies that allow for so-called pragmatic truth-values (von Fintel 2004). Testing whether these approaches are in line with speakers' actual judgments will inform the second set of hypotheses that we will be testing in the experiments.

(2) The Exhibition was visited by the king of France yesterday. (false)

The mentioned proposals will be discussed in Section 2. Afterwards, we turn to the review of some prior experimental work on this topic in Section 3 before moving on to the two central pieces of this contribution: the introduction of the *variance hypothesis* in Section 4 and the two experiments in Sections 5 and 6. For the variance hypothesis, we will consider the possibility that squeamishness is detectable not just introspectively but also experimentally by considering the consistency with which people respond to gappy experimental conditions. The hypothesis is that squeamish participants will show more uncertainty in their responses, which will be measurable via the most common index of variability, the standard deviation.[2] Throughout the two experiments presented afterwards, which also test truth-value intuitions directly via the employed scale, we will see evidence for this hypothesis, which, together with the other results, will suggest that uniqueness but not existence is an inference of the singular *the* that licenses a truth-value gap. All of this will be reviewed in greater detail in Section 7, in conjunction with the other hypothesis our data are informative about. Section 8 concludes.

2 Background

2.1 Russel and (Weak) Frege

The presuppositional status of the definite article remains controversial, and has been from the start. The three most relevant views are highlighted in (3) below.[3] While the first two are well known and have been

[1] In (1) and throughout I annotate truth-value intuitions reported in the literature right-dislocated and in parentheses. As is shown in (1), I will use # to indicate reports of presupposition failure; otherwise I will simply stick to the terms *true* and *false*.

[2] While the term *variance* is a technical term in statistics, I will use it in its non-technical sense throughout. The statistical parameter we use for the detection of presupposition failure is the standard deviation, never the variance (standard deviation squared). Unfortunately, 'standard deviation hypothesis' does not have quite the same ring to it.

[3] We will cast this description in terms of a partial function understanding of presuppositions. It is arguably more standard nowadays in research on presuppositions to assume a trivalent logic like Strong Kleene (Beaver & Krahmer 2001, George 2010, Fox 2013: e.g.). For our issues here, this choice is not consequential, as far as I can see. We will return to the question of partiality when turning to the predictions of Strawsonian squeamishness, where we will happen upon a different source of squeamishness in total logics.

the subject of much discussion, the third by Coppock & Beaver (2015) is more recent and deserves a more detailed discussion; also because it makes some more fine-grained predictions about the presuppositional status of singular definites that we will test in the experiments to come.

(3) The presupposition of *the P*

Russellian	$\varnothing$	−Existence; −Uniqueness; see Russell (1905)
Fregean	$\lvert P\rvert = 1$	+Existence; +Uniqueness; see Heim & Kratzer (1998) and Elbourne (2013)
Weak Fregean	$\lvert P\rvert \leqslant 1$	−Existence; +Uniqueness; see Coppock & Beaver (2015)

Before moving on to the analysis advanced in Coppock & Beaver (2015), let us briefly turn to the Russellian and Fregean positions indicated in (3). For our experimental purposes, these two views make the predictions indicated in (4a) and (4b) below. The Russellian view, now not generally supported, holds that existence and uniqueness are mere entailments of the definite descriptions, part of the assertion. The Fregean view, on the other hand, treats both existence and uniqueness as part of the presupposed content, not the truth-conditional content of the utterance. Presumably, it is fair to say that the Fregean position is the standard one in the linguistics literature, as it is also the one that is most commonly assumed in formal semantics textbooks (e.g., Heim & Kratzer 1998, Coppock & Champollion 2018).

(4) The king of France is bald.

a. *Presupposes nothing* (Russellian)

b. *Presupposes:* There is exactly one king of France. (Fregean)

The predictions that Russellians and Fregeans are committed to are straightforward: if either existence or uniqueness is violated, a Fregean semantics predicts presupposition failure, while a Russellian one predicts mere falsity. The third type of analysis mentioned in (3), the Weak Fregean approach from Coppock & Beaver (2015), leads to a more varied set of predictions than is not immediately apparent from the (3): while it is true that *the* does not add an existence presupposition according to Coppock & Beaver (2015), an existence presupposition of the utterance may still be achieved in some cases through the use of a type shifter. Before we turn to the implementational details, let us first have a look at their motivating examples.

Coppock & Beaver (2015) go through a number of examples to support their conclusion that singular *the* triggers a uniqueness but not an existence presupposition. In the interest of brevity, I will only walk through a few of these, and refer the reader to Coppock & Beaver's paper for a more comprehensive discussion. A central piece of evidence for them are so-called anti-uniqueness effects that arise when the definite description contains adjectival *only*. First, consider (5a) (Coppock & Beaver 2015: p. 384). Here, the interpretation suggests indeed that there is only a single author of *Waverley* and that this author is Scott. When embedded under a downward-entailing operator, however, here exemplified using negation, an existence presupposition associated with *the* suggests a reading according to which the Scott is not an author of *Waverley* at all. This is in conflict with the most natural reading of the sentence, namely that Scott is an author of *Waverley*, just not the only one. If instead we assume with Coppock & Beaver that *the* only triggers uniqueness, we arrive at the desired interpretation according to which *the* presupposes that there is at most one 'only author of *Waverley*'. Scott is not the individual x such that no person besides x is an author of *Waverley*. If we assume further than adjectival *only* presupposes that its individual argument satisfies its predicate argument,[4] then we additionally derive the entailment that Scott is an author of *Waverley*, and arrive at the desired interpretation according to which there is more than one author.

(5) a. Scott is the only author of *Waverley*.
⇝ There is exactly one author.

[4] Coppock & Beaver (2015) assume the following lexical entry for *only*:

(i) $\lambda P\,.\,\lambda x\colon P(x)\,.\,\forall y[x \neq y \rightarrow \neg P(y)]$

b. Scott is not the only author of *Waverley*.
   ⇝ There is more than one author.

While the existence presupposition can be made to disappear in various ways, uniqueness is much more stubborn, even in ignorance contexts (Coppock & Beaver 2015: p. 393). In (6a), the speaker seems to make the assumption that iguanas, who they are not sure have a heart at all—in violation of existence—, have at most one heart, and we detect felicity. Under the assumption that iguanas have multiple bones if they have any, on the other hand, (6b) is infelicitous. Coppock & Beaver interpret this contrast as suggesting that existence is not presupposed, while (weak) uniqueness is, which upon violation in (6b) triggers presupposition failure.[5]

(6) a. I don't know whether iguanas have hearts, but is that the heart?
    b. # I don't know whether iguanas have bones, but is that the bone?

As a control that the judgment in (6b) is actually due to presupposition failure, compare (7a) where we repeat the configuration from above but test the factivity presupposition of *know* instead. As in (6b), we observe the kind of infelicity we expect from a projective factive presupposition that iguanas have hearts clashing with the initial assertion of ignorance. By contrast, if we replace the definite with an indefinite, (7b), the utterance is unobjectionable, suggesting that uniqueness is indeed to blame.

(7) a. # I don't know whether iguanas have bones, but does Timmy know that they do?
    b. I don't know whether iguanas have bones, but is that a bone?

For another point of contrast, we will look at what has been called metalinguistic negation (Horn 1972), which allows (among other things) for at-issue, non-projective interpretations of presupposed content. To see this, note that (8a) is only felicitous with heavy emphasis on the definite, which in turn appears to allow for a denial of the uniqueness inference. By contrast, the variant of (8a) where we attempt to deny existence is not acceptable, contrary to our expectations regarding that prosodic profile and the assumption that existence is presupposed:

(8) a. It's not true that THE heart of the iguana exists, there are two.
    b. # It's not true that THE heart of the iguana exists, there isn't one.

As before, uniqueness, but not existence, seems to run parallel to what we find with the factive presupposition triggered by *know*:

(9) It is not true that Mary KNOWS that iguanas don't have bones, they do have them.

With this background in mind, we can now turn to the way absence of existence is implemented semantically in Coppock & Beaver (2015). Along the way, we will see more fine-grained predictions, including the reappearance of presupposed existence with argumental definites. The basic semantics for *the* that Coppock & Beaver (2015) give is listed below:

(10) $[\![\textbf{the}]\!] = \lambda P_{\langle e,t\rangle} : |P| \leqslant 1 \,.\, \lambda x_e \,.\, P(x)$

For singular definite in predicative uses, this derives the desired result of weak uniqueness, the presupposition that there be at most one individual that satisfies the description. In order to also capture, say, definites in subject position, Coppock & Beaver employ a range of type shifters from Partee (1986) that return argumental definites. They come in two flavors, determinate and indeterminate. The determinate type shifter IOTA, as the name suggests, introduces an existence presupposition using ι and returns an individual interpretation. The other option, indeterminate EX, instead shifts the predicate into an existential quantifier, which does not trigger any additional presuppositions. The two type shifters are defined as follows:

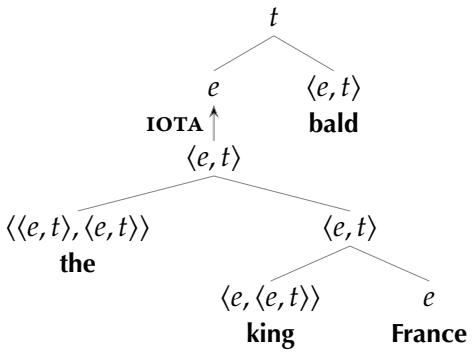
[5] Here, on may wonder whether the contrast between existence and uniqueness could be due to what is sometimes called hardness of the uniqueness presupposition (but softness of the existence one) such that some suspension mechanism applies and causes the alleged existence presupposition not to project (see, for example, Beaver & Krahmer 2001, Abusch 2002, 2010). We will return to the issue of suspension/local accommodation in detail later.

(11) a. $[\![\textsc{iota}]\!] = \lambda P_{\langle e,t\rangle} . \iota x[P(x)]$ determinate
b. $[\![\textsc{ex}]\!] = \lambda P_{\langle e,t\rangle} . \lambda Q_{\langle e,t\rangle} . \exists x[P(x) \wedge Q(x)]$ indeterminate

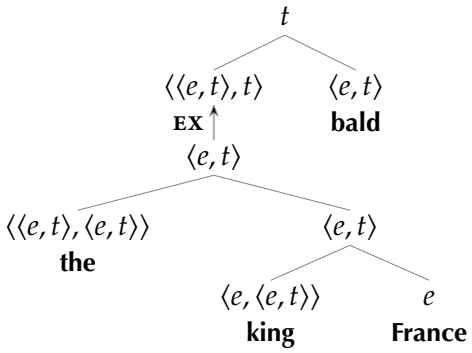
In this way, while predicative uses of *the* lead to utterances that presuppose weak uniqueness, argumental uses can be made to additionally presuppose existence or not, depending on the type shifter used to generate the appropriate type. The two LFs below illustrate this typal and ultimately semantic difference, with arrow-shaped (and appropriately labeled) branches indicating type-shifting.

(12) The king of France is bald.

a. *Presupposes:* There is exactly one king of France.



b. *Presupposes:* There is at most one king of France.



As for the choice between these two, Coppock & Beaver (2015) argue that IOTA should be preferred because it leads to a simpler output type, resulting in a general prioritization of Fregean (12a) over its weakly unique variant with EX. Only when the context indicates that a speaker might not take existence to be presupposed, as is possible for (13), should EX be used.

(13) Anna didn't give the only invited talk. (Coppock & Beaver 2015: p. 413)

a. There was more than one invited talk, one of which Anna gave. (anti-uniqueness reading)

b. There was exactly one invited talk, and Anna didn't give it. (determinate reading)

Overall then, Coppock & Beaver (2015) argue that *the* lexically only requires weak uniqueness to be entailed by the context. This is true for both the predicative case and indeterminate uses of argumental *the*. With the preferred determinate type shifter IOTA, argumental definite descriptions additionally presuppose existence, yielding the Fregean pattern. In addition, we expect a violation of (weak) uniqueness to lead to intuitions of presupposition failure no matter the semantic type.

To conclude this section, (14) contains a summary of the three approaches in terms of potential lexical entries for singular *the*, under the assumption of a predicative semantics to maintain consistency with Coppock & Beaver (2015). In what is to come, we will have a look at potential consequences of violations of existence and uniqueness. Rather than staying at the level of available patterns (14) suggests, we will see approaches that achieve classical truth values (or at least intuitions there) in spite of presupposition failure, further complicating the empirical predictions we will have to deal with experimentally.

(14) a. $[\![\textbf{the}_{\textbf{Russellian}}]\!]_{\langle\langle e,t\rangle,\langle e,t\rangle\rangle} = \lambda P_{\langle e,t\rangle} . \lambda x_e . P(x)$

b. $[\![\textbf{the}_{\textbf{Fregean}}]\!]_{\langle\langle e,t\rangle,\langle e,t\rangle\rangle} = \lambda P_{\langle e,t\rangle} : |P| = 1 . \lambda x_e . P(x)$

c. $[\![\textbf{the}_{\textbf{Weak Fregean}}]\!]_{\langle\langle e,t\rangle,\langle e,t\rangle\rangle} = \lambda P_{\langle e,t\rangle} : |P| \leqslant 1 . \lambda x_e . P(x)$

2.2 Existence is fickle

To start with the complications associated with singular definites, we do not have to venture very far, though we will be concerned mostly with the truth-value intuitions for utterances in contexts that violate existence. When world knowledge is involved, contexts that should not admit an utterance with an existence presupposition can be made to lead to truth-value judgments that are not indicative of presupposition failure. For an illustration, consider the case in (15): since our solar system does not contain a tenth planet, the definite description should result in a judgment of undefinedness. The absence of squeamishness may then be taken to suggest that existence is in fact not a presupposition of the definite determiner. An alternative, and perhaps not quite as radical, diagnosis is that there are cases where the violation of existence presuppositions may be charitably ignored in order to avoid the breakdown of the conversation (Geurts 2008: as suggested in). Especially in cases like (15), where the speaker is likely aware of the absence of a tenth planet, this appears to be a plausible explanation.

(15) Pluto is not the tenth planet from the sun. (true)

Yet another explanation for the apparent lack of presupposition failure above will become clearer with another example, (16) (from Strawson 1964). At face value, the minimal contrast appears to revolve around linear order. In (16a), the reported judgments are consistent with a view of singular definites as triggering existence presuppositions. The definite description in (16b), on the other hand, does not give rise to such intuitions of undefinedness, but rather to classical falsity—which, of course, begs an explanation on the lexical semantics we have been assuming thus far. The response in the literature, starting with Strawson (1964) himself, is that topicality, manipulated via linear order, is the deciding difference between the two. On this view, while the lexical contribution of the definite determiner is the same in both cases, information structure affects truth-value intuitions such that the violation of topical presuppositions leads to intuitions of genuine presupposition failure while non-topical presuppositions do not fall into a truth-value gap. That same explanation may then be applied to (15), where falsity follows from the violation of a non-topical presupposition.

(16) a. The king of France visited the Exhibition yesterday. (#)

b. The Exhibition was visited by the king of France yesterday. (false)

A common view in the literature has thus been to reduce judgment asymmetries of the type above to information structural categories such that topical expressions with presupposition failure lead to (intuitions of) undefinedness (Strawson 1950, 1964, Geurts 2008 a.o.).

According to Strawson, subjects and definites (as well as other referring expressions) are more likely to be topical. The manipulation of voice in (16) arguably exploits these tendencies and achieves a changed topic compared to the minimal variant, and, as predicted by this account, a changed truth-value judgment. In addition, Strawson (1964: p. 95) argues that sentence-initial material is likely to be topical, which is what we will exploit in the experiments to come.[6]

[6] Partee (1996) argues for a view of the conditions that allow for local accommodation that is similar to Strawson's presented above. Partee argues that, following work by Hajičová (1984), in the general case, non-projection interpretations of originally presupposed meaning components are only available if their trigger occurs in a non-topic position. If a definite description in our case is part of the topic, local accommodation is unavailable, and the presupposition projects.

Since this approach is primarily concerned with local accommodation in the sense of Heim (1982), which is not easily implemented in non-embedded cases which form the basis for our experimental endeavors later, I will unfortunately have to leave a detailed presentation to other work. In the context of our current project and allowing for an operator-based definition of local accommodation like in Strong Kleene (Beaver & Krahmer 2001), the treatment in Partee (1996) makes the same predictions as Strawsonian topic dependency.

In seeming defiance of the prediction Strawson makes, the following data from Lasersohn (1993: p. 113) suggest that topicality is not the factor that determines whether presupposition failure leads to squeamishness. In the examples below, the empty definite descriptions are both sentence-initial and a subject, yet the reported judgments violate our expectations of undefinedness.

(17) a. The king of France is sitting in that chair. (false)
b. The king of France is knocking on the door. (false)
c. The king of France ate that sandwich. (false)

In discussion of these data, Geurts (2008), who endorses Strawson's treatment, considers the possibility that there is a general tendency for contextually salient elements to be preferentially interpreted as topics. Taken together with what Geurts identifies as another key factor of conversation, the principle of charity, this leads to the expectation that speakers avoid infelicity, and by extension undefinedness,[7] by choosing the topic that is most likely to allow for the discourse to continue. For (17) in particular, a charitable speaker would choose not to designate the king of France as the topic but rather the chair, the door, or the sandwich, respectively.

Amending the classical example with a bit of context, Strawson (1964) now diagnoses falsity. This prediction mirrors the one in Schoubye (2009), who does not take topicality to be the critical factor, but rather the assumed QUD—see Footnote 9.

(18) A: What examples, if any, are there of famous contemporary figures who are bald?
B: The king of France is bald. (false)

2.2.1 Belief-revision

For the topicality-based view, truth-value intuitions for utterances normally expected to suffer from presupposition failure reflect actual truth values output by the compositional process. For von Fintel (2004), in sharp contrast, the truth value generated by the compositional process is not necessarily identical to the truth value that speakers assign to the utterance. He argues that even when presupposition failure occurs, a pragmatic filter may apply that yields a binary, though pragmatic truth value. Underlying this view is the assumption that speakers are able to verify what the truth value of the utterance would be if the presupposition had been satisfied in the context. If in the course of this revisionist process the speaker finds that the utterance would have been assigned false even with the presupposition satisfied, the pragmatic truth value is FALSE. If, on the the other hand, the truth value of the utterance cannot be determined even in this imaginary revised context, no pragmatic adjustment of the truth value occurs, and presupposition failure is retained.

Informally, the example below illustrates the verification-based system that von Fintel (2004) argues for (as an elaboration of the ideas in Lasersohn 1993).[8] The first example is semantically undefined and remains so even in when we pretend that there is a king of France: whether the pretend-king owns a pen is not something that can be settled by assuming that the presupposition holds in the current context. By contrast, the second example allows for a much easier verification once the presupposition is assumed to hold: since the pen is present in the context of utterance, it can be checked who it's owner is, and this

One caveat inherent to this approach lies with other potential applicability conditions of local accommodation. Relevant for our case is the intuition that only soft presuppositions are eligible for local accommodation while hard triggers resist this interpretation (Abrusán 2016, Chen, Thalmann & Antomo 2022, Thalmann & Matticchio (to appear)). If uniqueness and existence are both soft presuppositions, we expect them to show comparable amounts of fickleness. We will discuss whether the assumption of softness and the prediction of such an account are plausible in view of our experimental results in Section 7.

[7] We will return to the question of what motivates this avoidance of undefinedness later, though our discussion will center around the principle called Stalnaker's Bridge (von Fintel 2008, Fox 2013).

[8] There is a fundamental similarity in the two approaches argued for by Lasersohn (1993) and von Fintel (2004), which, for reasons of space, I will not have the opportunity to highlight. I do want to stress, however, that even in their notion of revision the two systems share a striking similarity. This is to say that even though the implementational details differ—for Lasersohn (1993) truth is relative to data sets—, I take the empirical predictions of the two approaches to be identical when it comes to the experimental results I will present later.

ownership will be found to be false, even if we revise our contextual knowledge to include a king of France. Hence, this second examples is assigned a pragmatic truth value of FALSE.

(19) a. The king of France owns a pen. (#)
b. The king of France owns this pen. (false)

Actual verification is not needed, however. According to von Fintel (2004), the mere possibility of verification is sufficient to yield a pragmatic truth value. Consider the example below, which is said to be judged as false (von Fintel 2004: p. 284). Here, we need not actually inspect or even be in a position to inspect Australia's state visit schedules; it is enough that we can check the properties of Australia in principle to show that there is no king of France there this week. What is needed is a contextually salient entity that could in principle falsify the statement, even if it does not actually do so—*a pen* and *bald* are not sufficient. Since revising our beliefs about the world to entail the existence of a king of France would not lead to a situation where the king of France is on a state visit to Australia, the utterance yields a pragmatic truth value of FALSE.

(20) *Context*: You and I have no reliable knowledge about Australia's current events.
The king of France is on a state visit to Australia this week. (false)

In contrast to the cases above, the classical example below is unaffected by the possibility for pragmatic adjustment: for the bald king of France, a minimal revision of our contextual knowledge to remove the presupposition failure will not have anything to say about the king's hairedness. In some compatible worlds, he is bald, in others he is not, yielding undefinedness overall and leaving our judgment unchanged.

(21) The king of France is bald. (#)

To capture these intuitions more formally, von Fintel (2004) proposes a notion of belief revision, which allows for the consideration of information states where no presupposition failure obtains, (22), and an operation to reject (pragmatically judge as FALSE) those utterances which are evaluated as false in these revised information states, (23). Together these deliver the pragmatic truth values that we saw above.

(22) Conversational revision (von Fintel 2004: p. 286)
a. Remove $\neg\pi$ from D (a body of information).
b. Remove any proposition from D that is incompatible with π.
c. Remove any proposition from D that was in D just because $\neg\pi$ was in D, unless it could be shown to be true by examining the intrinsic properties of a contextually salient entity.
d. Add π to D.
e. Close under logical consequence.

(23) Rejection (von Fintel 2004: p. 281)
Reject a sentence ϕ (with presupposition π) as FALSE with respect to a body of information D iff for all worlds w compatible with $rev_\pi(D)$: $[\![\phi]\!](w) = 0$

With this mechanism of generating pragmatic, binary truth-values from sentences that suffer from semantic presupposition failure, von Fintel (2004) gives us a handle on scenarios where the context does not entail the presupposition of an utterance obtains but where this turns out not to be catastrophic.[9] Lastly,

[9] Schoubye (2009) notices that mental states on von Fintel's view should not be good enough verification footholds. Conservative belief revision should not tell us anything about the desires and beliefs of the King of France. Yet, it is fairly easy to construct sentences where falsity obtains. There is no contextually salient entity whose properties we could inspect for verification—unless we weaken verification in such a way as to be able to inspect the mental states of non-existent entities.

(ii) The King of France wants to steal your car. (false)

In remedy of this, Schoubye (2009) suggests that the driving force behind the non-squeamish truth-value intuitions are essentially due to the interplay between (potentially implicit) questions under discussion with answers to these questions suffering

this formulation explains the contrast between the classical baldness example, (21) and (24) below: since the cases in (24) both provide verification footholds, namely the set of bald people in this world, we can reject and judge these cases as FALSE despite the semantic presupposition failure. As von Fintel (2004) argues, his revision analysis improves over a purely topic-based view of non-catastrophic presupposition failure since it does not require the presupposition to be topical to derive falsity. Further, the data from Lasersohn (1993) in (17), which are somewhat cumbersome to explain based on Strawsonian topicality, fall into place much more naturally when one assumes a belief-revisionist perspective: in all three cases, a verification foothold is provided, and a pragmatic truth value is derivable. Similarly cumbersome is (24) where topicality likely predicts a contrast such that the first, ostensibly topical, case comes out undefined while the non-topical variant should be judged as false, contrary to the judgments reported by von Fintel (2004: p. 286).

(24) a. The king of France is one of the bald people in this world. (false)
b. Among the bald people in this world is the king of France. (false)

Before concluding this section, it bears mentioning that while von Fintel (2004) does not explicitly discuss the local accommodation implementation of the intuition championed by Strawson (1964), he does argue against approaches that make use of this operation in general. In support of this conclusion, he relies on the following contrasts and argues for a strict separation of semantic truth value and truth-value intuitions.

(25) A′: The Exhibition was visited by the king of France yesterday. (false)
B′: Hey, wait a minute—I had no idea France was still a monarchy.

The 'Hey, wait a minute diagnostic' response is acceptable only when it targets non-at-issue meaning components; when it targets at-issue meanings, its use results in infelicity. One prominent view of presuppositions is that it is their non-at-issueness which makes them project (for discussion and experimental evidence, see Chen, Thalmann & Antomo 2022). Given that local accommodation stops projection, a common view is that local accommodation makes presuppositions at-issue. Now, since the responses in (25) are acceptable, von Fintel (2004) argues that the existence presupposition is not at-issue. However, if the truth-value intuition of falsity in either case were the result of local accommodation, the 'Hey, wait a minute diagnostic' responses should be infelicitous, contrary to fact. From this, von Fintel (2004) concludes that the falsity detected here is not the result of local accommodation, but rather the result of a verification-based system that yields pragmatic truth values (while leaving the semantic presupposition and its non-at-issueness unaffected).

von Fintel (2004) only discusses the interplay of belief-revision and the rejection mechanism in the context of existence presuppositions, but nothing prevents us from applying (22) and (23) to uniqueness presuppositions as well. Though the predictions for uniqueness presuppositions, especially in the context of our experiment, are less clear, we will discuss the implications of this verification-based system for uniqueness presuppositions in the experimental section.

While von Fintel does not explicitly discuss the experimental implications of his verification-based system, one plausible expectation is that truth values that are pragmatically derived should pattern together with other pragmatic enrichments, such as scalar implicatures. Supporting evidence for this view should then be a correspondence between judgments of falsity in the face of semantic undefinedness and the judgments for violations of scalar implicatures. In addition, since pragmatic processes are generally assumed

from presupposition failure. Schoubye argues that speakers essentially group false answers and answers with presupposition failures together: since someone asking a question chiefly cares about the resolution of the posed QUD, all non-true answers have the same status. On the other hand, answers that no matter their actual truth value could not resolve the QUD are judged as uncooperative and thus trigger squeamishness.

As far as I can see, for the purposes of my experiments, the QUD-based approach in Schoubye (2009) makes predictions that are largely in line with the predictions in Strawson (1964) on the assumption that permutations of linear order affect the assumed QUD by manipulating focus potentials (as assumed by Schoubye 2009: p. 606).

to be optional, we also expect that the pragmatic truth values should be associated with more heterogeneity than truth values where no pragmatic adjustment takes place. We will return to this when discussing the experimental predictions.

3 Previous work

To our knowledge, the only study that investigated to what extent the violation of topical/non-verifiable presuppositions differs from non-topical/verifiable counterparts in terms of squeamishness is Abrusán & Szendrői (2013). In their study, participants were told that they were to participate in a trivia quiz which tested their world knowledge. This setup allowed for the inclusion of actual examples from the literature as experimental stimuli. (26) shows the instructions and sample materials from their study.

(26) *If you think a statement is true, you should click on the 'TRUE' button. If you think a statement is false, you should click on the 'FALSE' button. Sometimes, it may happen that you cannot decide. In those cases, you should click on the 'CAN'T SAY' button. …There is no right or wrong answer!*
(Abrusán & Szendrői 2013)

The king of France is (not) bald.
The king of France is (not) on a state visit to Australia this week.
The king of France is (not) married to Carla Bruni.
The king of France, he was (not) invited to have dinner with Sarkozy.
Sarkozy, he was (not) invited to have dinner with the king of France.

FALSE CAN'T SAY TRUE

Abrusán & Szendrői (2013) find no effects at all in the positive versions of the items, with rejection rates remaining low throughout. Instead, participants overwhelmingly responded to the different non-negated conditions in (26) using the FALSE option. Even in the negative versions, CAN'T SAY responses were rarely used, and the overall patterns are hard to interpret. For example, the negated baseline condition 'The king of France is not bald' was predicted by none of the accounts under investigation (including von Fintel 2004) to trigger anything but squeamishness, yet participants only rejected this condition 33.9% of the time. Now, it is possible that participants did not uniformly use the CAN'T SAY option to communicate squeamishness, but this is hard to establish in this experiment. Additionally, the FALSE response option, which was predicted by no account, was chosen 44.9% of the time, making it the most popular response for the baseline condition. Since the mapping from predicted truth value to experimentally selected response options seems to have been noisy even for the control, drawing conclusions on the basis of divergence from that control is fraught with difficulty.

Regarding the goal of testing the influence of topicality, even though well-discussed examples from the literature were used in the experimental items, topic status is difficult to diagnose in a task that is set up as a quiz where everything could easily be read as all-new.

In addition, the instructions for the experiment seem contradictory overall: for one, participants are tested on their world knowledge in a trivia quiz—a setting where wrong answers typically cause you not to win (and options like CAN'T SAY seem inappropriate). On the other hand, participants received the instruction that there were no right or wrong answers. Juxtaposed with the absence of explicit controls for squeamishness, the results presented in Abrusán & Szendrői (2013) require further exploration using other experiments in order to address the questions of the study.

4 The variance hypothesis

Before moving on to the experiments that were designed to test the hypothesis associated with the accounts we looked at above, I want to highlight a hypothesis that has so far not received any attention in the experimental literature on presupposed meanings.

There is a portion of the intuition behind, and to some extent the theoretical modeling of, presuppositions that is often ignored in experimental studies dealing with presupposition failure. It bears repeating that this core intuition is perhaps best described with the term 'squeamishness' in Strawson (1964) and others, and it is meant to capture the uneasiness that speakers confronted with (certain kinds of) utterances suffering from presupposition failure display. Essentially, different from sentences that are truth-evaluable, encountering presupposition failure leads to a breakdown of communication and may leave interlocutors unable to respond without taking the time to highlight the occurrence of this breakdown.[10] Why presupposition failure amounts to a crash of the ongoing conversation is easily appreciated in semantic systems where presuppositions are encoded as definedness conditions (as is the case in, for example, Heim & Kratzer 1998). An utterance whose meaning cannot be computed because it contains a (partial) function that is undefined for its argument is naturally expected to give the interlocutors pause since truth-evaluability is simply not possibility. Rather than furthering the conversational goals, in these systems utterances suffering from presupposition failure lead to unevaluable discourse move.

So in a semantics where presupposition failure essentially mimics the uninterpretability following from type mismatches, squeamishness is a plausible consequence. Non-binary logics, however, are less straightforwardly interpreted this way because they generally conceive of presupposition failure as receiving a defined truth value, just one that is distinct from truth or falsity, e.g., with Strong Kleene (for an overview, see Beaver & Krahmer 2001). In these total-function systems, it is not immediately obvious why a third value assigned to utterances with presuppositions in a context that does not support those presuppositions should engender larger amounts of uncertainty if the utterance is not classified as semantically deficient. After all, from the perspective of these logics where all functions are total, classical truth values are in principle on par with non-classical ones; it's just that the mapping is a different one. Of course, this initial hurdle is not insurmountable, and there are implementations of three-valued logics in the literature that are amenable to a squeamishness interpretation. The clearest example of this type is Fox (2013), who adopts Stalnaker's Bridge (von Fintel 2008) in order to derive a pragmatic presupposition from the total functions that Strong Kleene delivers, (27).

(27) **Stalnaker's Bridge** (von Fintel 2008, Fox 2013)
A truth-denoting sentence S is assertable given a context set C only if $\forall w \in C[[\![\mathbf{S}]\!](w) = 1 \vee [\![\mathbf{S}]\!](w) = 0]$

What I tentatively suggest here is that Stalnaker's Bridge might also give us a handle on squeamishness, if not in a strictly semantic but rather a pragmatic fashion. That is, following Stalnaker's Bridge, which von Fintel (2008) and Fox (2013) argue for on independent grounds, we expect that a sentence that receives a non-classical truth value is deemed not assertable. This, in turn, gives us a handle on squeamishness: a non-assertable utterance that nonetheless is asserted is reasonably expected to put interlocutors in a position of conversational confusion, and thus, I argue, as displaying squeamishness. This gives us a linking hypothesis both for systems where presuppositions are partial and those where presuppositions represent total functions via the route of failed conversational turns.

It is precisely the effects of this breakdown that, as I want to advance in the present paper, should be measurable in (certain) experimental settings. While squeamishness may be related to longer reaction times, I want to pursue the hypothesis that it should be detectable in offline measures like ratings as well, at least when multiple trials with presupposition failure are considered, as will be the case in the experiments to be presented. Consistent with the description above, I take it that if squeamishness is induced by presupposition failure, it should lead to rating behaviors plagued by uncertainty. That is, over the course of the experiment, intuitions of squeamishness should elicit response patterns that are highly variable relative to sets of trials where no catastrophic presupposition failure obtains.

[10] This can also seen in the rather roundabout way that disagreements about presupposed meanings have to be signaled. Take for example the 'Hey, wait a minute' test (von Fintel 2004), which is odd when addressed at asserted content:

(iii) A: The university chaplain is bald.
B: Hey, wait a minute! I didn't know our university had a chaplain!
B′: # Hey, wait a minute! I didn't know the university chaplain is bald.

On this view, one way to detect (intuitions of) presupposition failure experimentally is to look at the standard deviations associated with the condition means: all else being equal, undefinedness, or rather the squeamishness triggered by it, should elicit larger uncertainty, which in turn can be measured through larger standard deviations. More precisely, this hypothesis predicts that conditions without presupposition failure should show a more strongly clustered rating distribution around the mean, i.e., lower standard deviations. Conversely, a more widely spread-out pattern, as measured by larger standard deviations, is expected for conditions where there is a presupposition that is not entailed by the context set and, to the extent that this is possible, which is not subject to mechanisms that nevertheless yield a binary truth value.

At this stage of the discussion, it is important to clarify that presupposition failure is unlikely to be the only modulator of standard deviations. Falsity may, for example, lead to higher standard deviations simply by virtue of being more challenging cognitively. In addition, if interpreted as an index of uncertainty, other features immediately spring to mind as drivers of larger variance. Among these are a wide variety of linguistic and extra-linguistic properties such as, for example, marked lexical material or syntax, complexity differences in visual stimuli, and the presence of pragmatic inferences. Of course, not all of them are active for all participants, and not all of them are knowable a priori. Hence, we will test whether increased standard deviations are interpretable as causally related to presupposition failure, just like with the means we normally tend to focus on.

Standard statistical models in linguistics and other experimental disciplines normally assume that the clustering around the means we estimate is the same between different experimental conditions. If a model makes this assumption, then the standard deviation between conditions is simply pooled and estimated for the entire data set at once, such that each mean in the model is associated with the same standard deviation estimate. Data where this procedure is appropriate (because all conditions display roughly equal amounts of spread) are what is called homoscedastic. In other words, standard models tend to assume homoscedasticity. By contrast, data sets where different conditions feature distinct amounts of variance around the estimated means are called heteroscedastic, and applying a model that assumes homoscedasticity to these data leads to a pooled calculation of the standard deviations, and thus to an estimate for the standard deviation that is a poor reflection of either case.

As an example, consider the simulated data in Figure 1, Panel A, where we have a homoscedastic data set with two conditions x_1 and x_2 on the left, and a heteroscedastic one on the right, as indicated using the shaded areas around the means.[11] In Figure 1, Panel B we see what a standard model may estimate for this case. Ignoring the other model parameters, we see a parameter called *Sigma* that represents the standard deviation associated with each condition on both the left and the right. The fact that both estimates are roughly the same despite the quite remarkable visual difference in the Panel A between the two data sets is instructive for the potential consequences that the assumption of homoscedasticity may have. Assuming that the *Sigma* estimate on the left is reliable, it is immediately obvious that the same parameter for the other data set is not a good indicator of either x_1 or x_2: x_1 on the right looks more strongly clustered than the two conditions on the left, and right x_2 appears to be more spread out.[12]

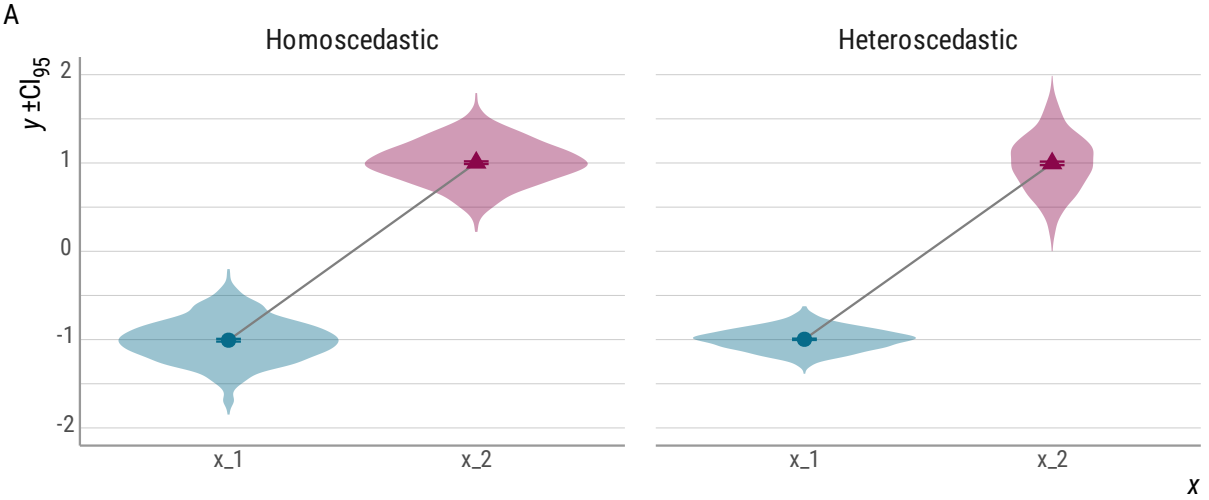
With these potential complexities in mind, I argue that relying on a model with the homoscedasticity assumption is a less parsimonious choice relative to a model where no such assumption is made and where standard deviations are estimated on a by-parameter basis. Since we cannot, in the absence of information about our data, know beforehand whether we are likely to find differences in by-condition standard deviations, adopting a model architecture with specific limitations on standard deviations risks faulty hypothesis evaluation.

The adequacy question of applying homoscedasticity-assuming models to heteroscedastic data notwithstanding, Figure 1, Panel B also shows that standard models will not do if our modeling goal also includes

[11] Notice here that due the amount of data that went into each mean represented in Figure 1, the confidence intervals around these means are not really indicative of the substantial variance differences in the heteroscedastic scenario.

[12] A violation of this assumption, i.e., pooling unequal variances, does not invalidate the model in itself, but it does lead to potentially biased inferences. More concretely, the mean estimates from heteroscedastic data are reliable but the inference that we base on mean differences is biased because the unsystematic variance is incorrectly assumed not to vary. Since the (residual) standard deviation enters into the calculation of, say, p-values, heterogeneity of variance impacts our conclusions if we use p-values as a criterion for evaluating hypotheses.

A



B

Standard models

	Homoscedastic data			Heteroscedastic data		
Parameter	Estimate	HDI_low	HDI_high	Estimate	HDI_low	HDI_high
Intercept	-1.0067	-1.0215	-0.9919	-0.9965	-1.0115	-0.9805
x_2	1.0036	0.9883	1.0189	0.9969	0.9823	1.0119
Sigma	0.2475	0.2403	0.2551	0.2473	0.2400	0.2549

C

Distributional models

	Homoscedastic data			Heteroscedastic data		
Parameter	Estimate	HDI_low	HDI_high	Estimate	HDI_low	HDI_high
Intercept	-1.0068	-1.0230	-0.9920	-0.9965	-1.0039	-0.9885
x_2	1.0038	0.9882	1.0187	0.9970	0.9780	1.0185
Sigma_x_1	0.2495	0.2393	0.2609	0.1266	0.1214	0.1318
Sigma_x_2	0.2456	0.2345	0.2566	0.3263	0.3113	0.3393

Figure 1: Illustration of the distributional model architecture for the investigation of the variance hypothesis for squeamishness following presupposition failure. *A*: The two data sets that will be compared, with the shaded areas indicating the raw rating distribution. *B*: The modeling results from standard, homoscedasticity-assuming models. *C*: Estimates and uncertainty intervals from distributional models, which do not assume homoscedasticity. Highest density intervals (HDIs) indicate the most credible values of the distribution that contain a specified proportion (95%) of the posterior probability.

the investigation of standard deviations as an index of squeamishness. If squeamishness is associated with presupposition failure, and if squeamishness impacts speakers' certainty, we have reasons to believe that presupposition failure will lead to an increase in standard deviation. But in order to compare standard deviations between conditions with and without presupposition failure, we need separate estimates, which provides us with another reason not to rely on models that architecturally assume homogeneity of variance. In remedy, we will instead switch from the standard model architecture to one where homoscedasticity is not assumed and where the model parameters include by-condition standard deviation estimates. What this achieves is shown in Figure 1, Panel C under the header of *Distributional models*, the name for models that can estimate multiple parameters of a distribution. On the left, we see two *Sigma* parameters with roughly identical estimates, on the right we see that these two parameters differ in their estimates. Encouragingly, the larger estimate on the right, the one associated with the x_2 condition, corresponds to the condition that shows less clustering visually in Panel A.

All in all, both of the reasons outlined above, the question of adequacy and our specific modeling goals, motivate the departure from the standard modeling procedure and the move to a richer architecture. For the purposes of our experiments, this means that the models we use will be a so-called distributional

one that estimates multiple parameters of a distribution. This move will then allow us to see whether it is indeed the case that Strawsonian squeamishness is a detectable feature of experimental responses via increases in standard deviations, as the leading intuition seems to predict.

5 Experiment 1

The first experiment, which was carried out in German and programmed using PsychoPy (Peirce et al. 2019), was designed to test a number of hypotheses associated with the (alleged) presuppositions of singular definite descriptions. Since the different predictions essentially all vary with respect to the truth value that is predicted, a natural choice for an experimental paradigm was one where responses correspond to truth-value intuitions as directly as possible. Abrusán & Szendrői (2013) already pursued this option but their results were hard to interpret at times, possibly also because the labels for these response options are not easy to determine. Luckily, in the meantime, Križ & Chemla (2015) in their experiments on homogeneity inferences licensed by definite plurals ran a number of experiments with the goal of finding labels that allow participants to communicate their gappy judgments. Ultimately, Križ & Chemla (2015) suggest the labels in (28) in a three-alternative forced choice task.

(28) *completely false*, *neither completely true nor completely false*, *completely true*

Since one of the main goals was detecting truth-value gaps and potential differences between gappy conditions, say because some participants apply a von Fintellian revision procedure, we decided to modify their experimental paradigm to potentially achieve higher discriminatory precision and more potential for variance (Sun, Schmidt & Henry 2025). Instead of employing a forced-choice paradigm, we made use of a continuous scale, labelled at the extremes and in the middle using the labels suggested in Križ & Chemla (2015). This led to the following scale:

(29) (completely false) (neither completely true nor completely false) (completely true)

With this continuous, trivalent truth-value judgment task, we decided to manipulate a number of features that were claimed to be operative for truth-value intuitions in the literature. In order to evaluate the accounts based on verification like Lasersohn (1993) and von Fintel (2004), we decided to use visual stimuli to manipulate the context in which participants evaluated the linguistic material. Ultimately, we settled on images showing 3×3 matrix-like displays filled partially with symbols, comparable to what was used in Križ & Chemla (2015). This also allowed us to avoid potential confounds relating to world knowledge as the driver of presupposition failure (cf. Abrusán & Szendrői 2013).

Next, to investigate the predictions in Strawson (1964) and Schoubye (2009), we decided to vary linear order as a way to manipulate topicality/inferred QUDs when discussing positional relationships between shapes/symbols, (30). While this is certainly an imperfect manipulation, it is a reasonable first step to test the hypothesis that presupposition failure may either be catastrophic or non-catastrophic depending on features of the context. No matter whether topicality is the relevant factor or whether the QUD mediates the difference, the expectation is that the two share a common core in cases where no overt question is given but stimuli are differentiated by linear order between conditions.

(30) a. The square is to the left of the triangle. (subject-initial)
b. To the left of the triangle is the square. (subject-final)

Finally, we decided to use as controls not only conditions where no presupposition failure obtains, i.e., classically true and false scenarios, but also conditions where we are reasonably sure from other experiments that participants will be able to detect a truth-value gap. The obvious choice for this condition again lies in Križ & Chemla (2015): definite plurals and the homogeneity inference they trigger, which were validated as detectable in an experimental paradigm that is extremely similar to one employed here. This choice also allowed for a more detailed investigation of the presuppositionality predictions in Coppock & Beaver (2015) and the proposals in Russell (1905) and Frege (1997), because we could compare

violations of existence and uniqueness to a syntactically related but semantically (plausibly) independent truth-value gap.

Below, we will walk through the experiment in detail.

5.1 Design and materials

In the design below we list the `true` and `false` scenarios as part of the critical items, for completeness sake, but the in terms of the experimental design, these were counted as part of the control conditions, since these scenarios were only used to check participants's performance.[13]

(31) $2 \times 2(\times 3)$-Design (all within-items and within-participants)

a. NUMBER: DP_{SG} vs./ DP_{PL}

b. STRUCTURE: subject-initial vs./ subject-final

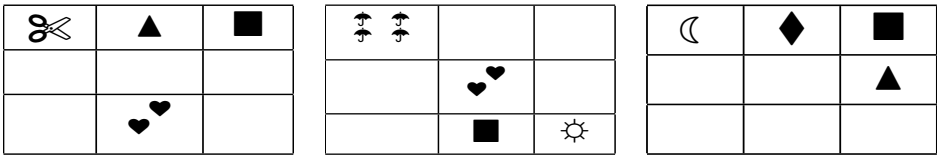
c. SCENARIO: `true` vs./ `false` vs./ `undefined`

In (32) below, we show a sample item in all of the conditions shown in (31). Both the shapes for the subject and the relatum varied throughout the items, as did the positional relations (*über* 'above', *unter* 'below', *links neben* 'to the left of', and *rechts neben* 'to the right of'). Of note is especially the inclusion of *direkt* 'directly' in all of the positional predicates, which was supposed to prevent unintended truth/falsity in configurations where shapes were not immediately adjacent. We will return to this design choice in the discussion section.

For the images, note in particular that the amount of symbols shown is always 4, no matter the condition. Existence violations are thus not predictable at a glance. Additionally, there was a mix of simplex and complex symbols in the images, the latter of which were necessary for the definite plurals. Participants saw six lexicalizations per condition, for a total of 72 items relating to (31).

(32) a. Das Dreieck steht direkt links neben dem Quadrat.
the triangle stands directly left next.to the square

b. Direkt links neben dem Quadrat steht das Dreieck.
directly left next.to the square stands the triangle



`true` `undefined` `false`

c. Die Dreiecke stehen direkt links neben dem Quadrat.
the triangles stand directly left next.to the square

d. Direkt links neben dem Quadrat stehen die Dreiecke.
directly left next.to the square stand the triangles

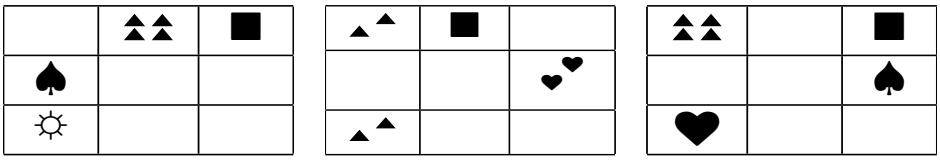
[13] While it may seem that the inclusion of the syntactic manipulation with the definite plurals was entirely for parallelity's sake, Križ (2015) discusses contrasts with homogeneity violations that resemble what Strawson (1964) and Schoubye (2009) mention for existence violations:

(iv) *Two of three boys and a girl carried the piano upstairs.* (Križ 2015: p. 13)

a. A: What happened?
B: Two boys carried the piano upstairs. (#)

b. A′: Who carried the piano upstairs?
B′: Two boys carried the piano upstairs. (false)

Though this is certainly not a focus of the present study, the word order manipulation does allow us to check whether the introspective judgments receive experimental support in the present paradigm.



`true` `undefined` `false`

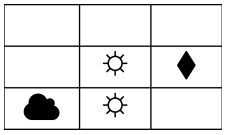
As is evident from the sample item above, violations of uniqueness were not included in what is labeled critical items above, which only included existence and homogeneity. Uniqueness violations were included in the control items. Just like the items above, the control items all came in two syntactic variants, subject initial and subject final, though all other manipulations represent distinct filler conditions. Participants saw each filler condition 6 times, 3 times each per syntactic variant, for a total of 48 fillers relating to the list below.

The filler conditions included violations of uniqueness, (33). In addition, we tested indefinite variants of the `false` and `undefined` scenarios, (34) and (35). Further, we included controls where the relatum symbols were positioned in such a way that a true sentence was impossible, both with and without an empty description, (36) and (37). Finally, since some of the accounts, especially Lasersohn (1993) and von Fintel (2004) assume a kind of pragmatic adjustment of the semantically derived truth values, we also included a condition which may allow for some degree of comparison in this regard, namely scalar implicatures. The hope was that, since scalar implicatures arguably represent an optional pragmatic process, there should be some similarity with these conditions and presupposition violations with pragmatic truth values; either in terms of the means or maybe with respect to the standard deviations. Examples for these conditions are in (38) and (39), with a true and a false implicature, respectively.

(33) `uniquenessviol`

a. Die Sonne steht direkt links neben der Wolke.
the sun stands directly left next.to the cloud

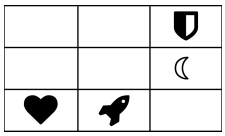
b. Direkt links neben der Wolke steht die Sonne.
directly right next.to the cloud stands the sun



(34) `indeffalse`

a. Ein Schild steht direkt links neben der Rakete.
a sign stands directly left next.to the rocket

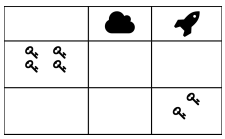
b. Direkt links neben der Rakete steht ein Schild.
directly right next.to the rocket stands a sign



(35) `indefundef`

a. Eine Sonne steht direkt unter der Wolke.
a sun stands directly under the cloud

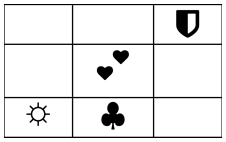
b. Direkt unter der Wolke steht eine Sonne.
directly under the cloud stands a sun



(36) `imposfalse`

a. Das Schild steht direkt unter der Sonne.
the sign stands directly under the sun

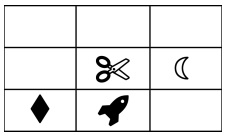
b. Direkt unter der Sonne steht das Schild.
directly under the sun stands the sign



(37) `imposundef`

a. Die Sonne steht direkt rechts neben dem Halbmond.
the sun stands directly right next.to the half.moon

b. Direkt rechts neben dem Halbmond steht die Sonne.
directly right next.to the half.moon stands the sun



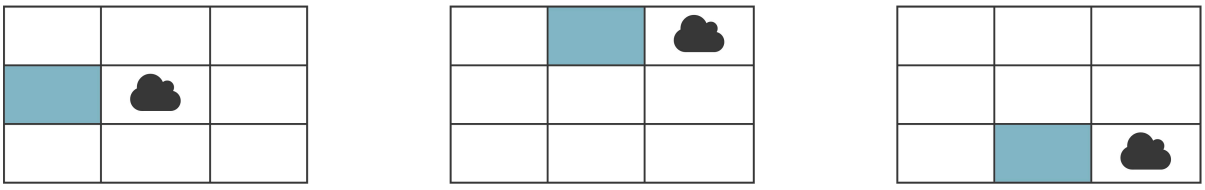
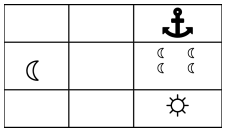


Figure 2: Series of images that participants were shown in the instruction portion of the experiment to explain the experimental task.

(38) `scalartrue`

a. Einige Halbmonde stehen direkt über der Sonne.
some half.moons stand directly above the sun

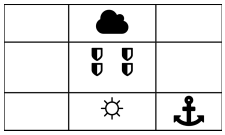
b. Direkt über der Sonne stehen einige Halbmonde.
directly above the sun stand some half.moons



(39) `scalarfalse`

a. Einige Schilder stehen direkt unter der Wolke.
some signs stand directly under the cloud

b. Direkt unter der Wolke stehen einige Schilder.
directly under the cloud stand some signs



5.2 Participants and procedure

We tested 24 participants (mean age 22.8 ± 2.6, 21 female identifying), who were recruited via university channels, in a laboratory setting at a computer. Participants received 7€ for their participation in the study, which took about 30 minutes on average.

At the beginning of the experiment, participants were informed that they would be presented with descriptions of shapes and symbols as well as accompanying images and that their task would be to rate using the scale in (29) whether the linguistic stimulus was true or false given the image. In addition to this general instruction, participants were shown the matrices in Figure 2 and were told that sentences with *directly left of the cloud* were to be verified using the highlighted cells in the images. As mentioned above, we return to the issue of *direkt* and the way it influenced the introductory portion of the experiment in the discussion section.

Afterwards, participants saw one true warm-up item where the reason for that judgment was explained on the basis of an image. While the neither-nor label of the scale was mentioned when the scale was introduced, it did not feature in any description. As a next step, participants started the experimental session with a randomized presentation of critical and control items. In these trials, participants saw both the text (on top) and the image (below) at the same time. At the very top of the screen, participants always saw the following instruction:

(40) Wie bewerten Sie den Satz in dem gegeben Szenario?
how rate you the sentence in the given scenario
'How do you rate the sentence in the given scenario?'

Before moving on to the predictions, let me briefly introduce the item-related data we collected. In addition to recording the final ratings for each item (i.e., the final position of the slider), we also recorded the real-time positions of the slider for each item from the very beginning for the entire duration until participants moved on to the next item. This was done not only to have access to reaction time data but to also be able to compute other behavioral indices that may be associated with presupposition-failure induced squeamishness. Of particular interest in this regard is the first motion of the slider, which may indicate whether participants were unsure about the rating they wanted to give in cases of presupposition failure. Another measure potentially associated with uncertainty is the number of direction changes of

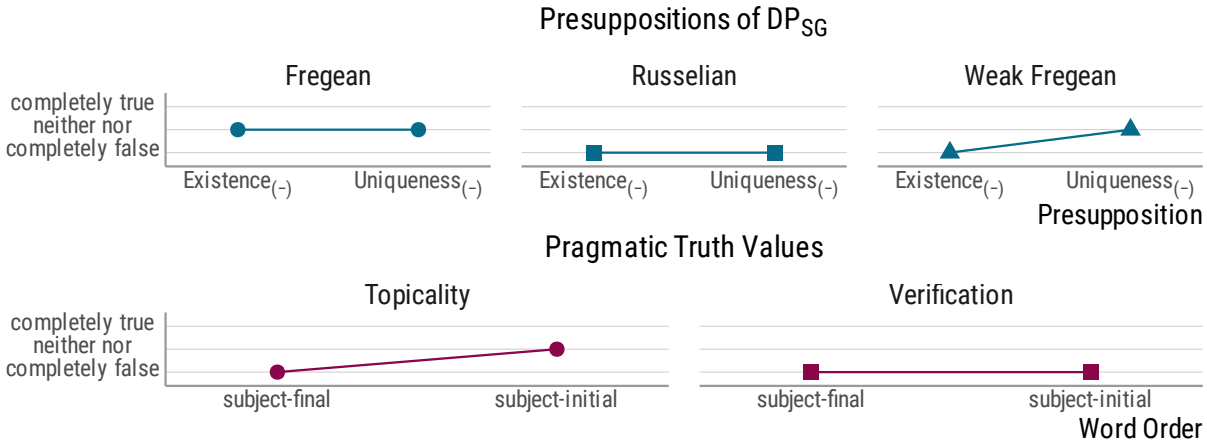


Figure 3: Overview of the various predictions.

the slider in the course of the rating process. Participants who change directions more often may do so because they change their decision while moving the slider, potentially because of squeamishness.

5.3 Hypotheses

Starting with the predictions, the most important, and method-validating one concerns the definite plurals. The present experiment can only be considered valid if we achieve a partial replication of Križ & Chemla (2015) such that we find three distinct rating patterns for the three scenarios with the DP$_{PL}$ items. In particular, we expect that the `undefined` scenario will be judged more scale-medially than the `false` and `true` scenarios, which should receive ratings at either extreme of the scale. The `false` and `true` scenarios with definite singular DPs are of course also expected to pattern this way.

Figure 3 shows the most important predictions graphically, with the top row indicating the predictions of the different approaches to the lexical semantics of singular *the* in (3). The bottom row highlights what we expect from the topicality view of squeamishness (Strawson 1964, Schoubye 2009) and the verification-based approaches (Lasersohn 1993, von Fintel 2004).

von Fintel (2004) predicts that pragmatic truth value adjustments should be possible in the UNDEFINED scenario because participants have full access to the matrix: even under the assumption that the symbol referred to by the definite description exists, the truth value of the sentence can be determined by looking at the properties of the contextually salient relatum. The inspection of the surrounding cells should then lead to the truth value of the sentence being determined as FALSE. Since the adjustment of the semantic truth value is pragmatic in nature, we do however expect both larger standard deviations and overall similarity between the UNDEFINED scenario in the critical materials and the SCALARFALSE condition in the controls.

It is not exactly clear what the predictions are for the DP$_{PL}$ items in the UNDEFINED scenario. Same for the condition where uniqueness is violated. In principle, we expect a revision mechanism like the one discussed in von Fintel (2004) not to be selective to particular presuppositions, but to allow for pragmatically adjusted truth values whenever the revision process achieves a result that does not lead to a breakdown of the conversation. If this is the case,

Turning away from the location-based hypotheses and moving on to standard deviations, we predict that the variance hypothesis holds. That is, when participants give squeamish judgments that indicate presupposition failure, we expect more dispersed judgments for those conditions. Note that we do not commit to a view according to which presupposition failure is the only possible reason for increased variances; we also expect that optional pragmatic processes should drive less concentrated rating clusters around the mean. The clearest example of this are the conditions with scalar implicatures, but we also expect similar increases in cases where truth values are optionally pragmatically altered (Lasersohn 1993, von Fintel 2004).

5.4 Statistical analysis

To analyze the experimental results, I fit a Bayesian mixed distributional linear mixed regression (using `brms` from Bürkner 2021 in R 4.5, R Core Team 2025) for the critical and the control items separately. The model ran with 8 chains with 10,000 iterations each, half of which were warm-up samples.[14] The core feature of distributional models is that they can model multiple distribution parameters simultaneously. In our case, we are interested both in the location parameter (the mean judgment) and the scale parameter (the standard deviation) of the distribution.

Model formulas for the critical items are shown in (41), with the location and scale parameters separated. As for fixed effects, the model contained the three factors from the design outlined in (31), as well as their interactions. For the location part of the model, the random effects were maximal with respect to our design. For the scale portion, (41b), no mixed effects were included to avoid estimation issues. Because standard deviations can never be negative, they were modeled on the log scale, though throughout our results, we will present back-transformed values. All predictors were sum-coded.

(41) a. $Y \sim$ TRIGGER * SCENARIO * STRUCTURE + (location)
(1 + TRIGGER * SCENARIO * STRUCTURE | PARTICIPANT) +
(1 + TRIGGER * SCENARIO * STRUCTURE | ITEM)

b. $\log \sigma \sim$ TRIGGER * SCENARIO * STRUCTURE (scale)

For the location portion of the model, the intercept and slope priors were set to be weakly informative, with a normal distribution centered on 0 and a standard deviation of .5 for the intercept and 1 for the slopes. To also reflect the outer limits of the scale, the intercept prior included upper and lower bounds at -2 and 2, respectively, and the slope priors at -4 and 4. The priors for the log standard deviations were similarly weakly informative normal distributions centered on 0, with a standard deviation of .5 for the intercept and .25 for the slopes.

The controls were analyzed using the model formula in (42), with COND representing each of the different conditions exemplified in (33) to (39). The priors mirrored those for the critical model. The COND factor was treatment coded.

(42) a. $Y \sim$ COND * STRUCTURE + (location)
(1 + COND * STRUCTURE | PARTICIPANT) +
(1 + COND * STRUCTURE | ITEM)

b. $\log \sigma \sim$ TRIGGER * COND * STRUCTURE (scale)

Rather than talking about the model coefficients, I will plot the posterior marginal distributions for the location and scale parameters. These were estimated using the `gather_emmeans_draws` function from the *emmeans* package (Lenth 2024). Conclusions will then be based on visual inspection of these posterior marginal parameter distributions, which represent probability densities of the estimates. In effect, using the associated credibility intervals, we can compare different conditions and decide whether the posterior marginal effects are comparable or whether they are more likely to represent a wholly different kind of meaning. In terms of our truth-value judgment task, we hope for a maximum of three kinds of posterior marginal means: those associated with true scenarios and false scenarios at either end of the scale, and means representing judgments of gappiness at some medial position. This last group, following the variance hypothesis, we also hope to come with large standard deviations There, too, we will rely on visual inspection of posterior marginal estimates.

5.5 Results

The ratings from Pavlovia originally ranged from 1 to 100, but were scaled to fall between -2 (completely false) and 2 (completely true) for interpretability reasons. Figures 4 and 5 show the ratings for the critical

[14] See Section 4 for a comparison of this model architecture to more standard linear mixed models.

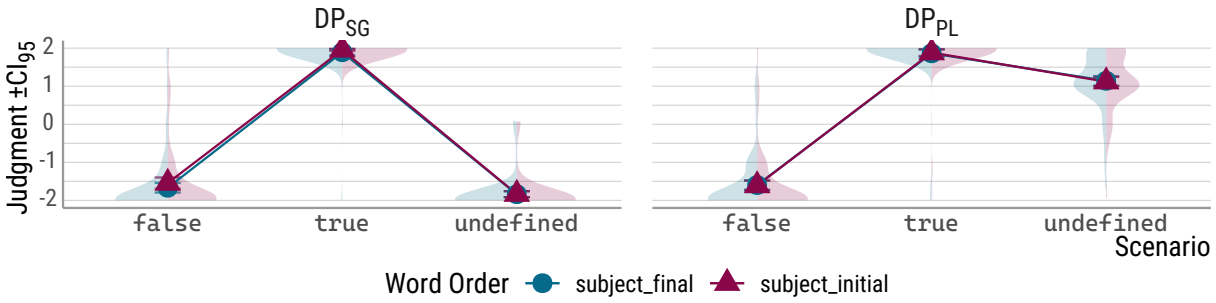


Figure 4: Ratings for the items in the critical portion of the experiment design. Shaded areas indicate the raw rating distributions that gave rise to the indicated mean judgments.

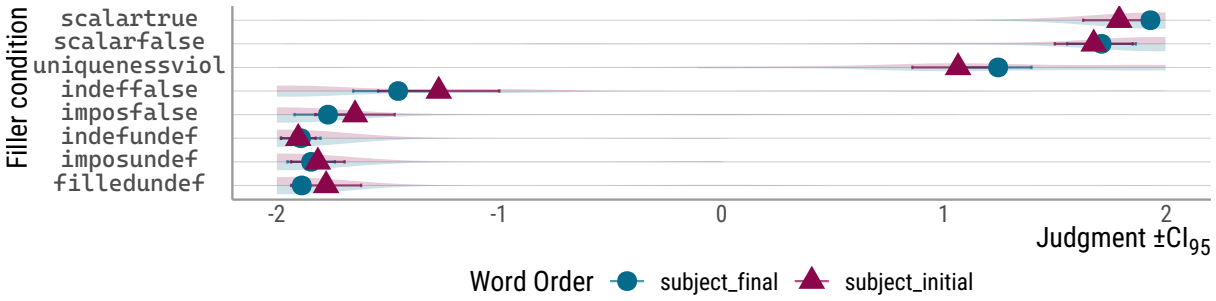


Figure 5: Ratings for the items in the control portion of the experiment design.

and the control portion, respectively. The additional measures were sought to investigate with the slider—reaction times, time to first motion, and direction changes—were not very informative and are presented in the appendix in Figure 12.

The coefficients for the fixed effects for both models are shown in the appendix, Section A.1.

For the calculation of the posterior marginal distributions with the *emmeans* package (**emeans**), I abstracted away from the STRUCTURE factor for both the critical model and the control model because it was not informative on its own and did not interact with the other factors in either case.

Figure 6 shows the posterior marginal means for the critical items and the controls. When inspecting the plots, it is important to keep in mind that the variance around the estimates represents uncertainty about the estimate for the mean and not the standard deviation around the mean. That is, a wider interval/distribution indicates that the location of the mean itself is more uncertain and not that the data around the mean are spread out more. Just like the raw results above, we see that model-adjusted means validate the method: the three scenarios with definite plurals show the expected pattern of a three-way distinction, such that the `true` and `false` scenarios received estimates at the extremes of the scale, while the condition where homogeneity was violated received a more medial judgment, if one that is closer to the *completely true* pole of the scale.

For the singular definites, we also find that the `true` and `false` receive posterior marginal means that align with our expectations. The `undefined` scenario, which tested violations of existence, was judged to be false, in parallel with the false controls for DP_{PL} and DP_{SG}. This rating pattern mirrors the posterior marginal means we find for the false control conditions in (34) to (37) as well, where we see an overlap of the two kinds of credibility intervals and hence little reason to suspect differences.

Two other control conditions are worth highlighting: the uniqueness violation condition, much like the homogeneity violation in critical items, has a posterior marginal location estimate that differs from both true and false controls. The condition where a scalar implicature was violated, on the other hand, is slightly more complex in that the estimate for it is so noisy that a difference/similarity between it and gappy conditions, like uniqueness and homogeneity violations, cannot easily be established. At the same time, it is not clear whether we should assume similarity between violations of scalar implicatures and true controls as well as non-violated scalar implicatures.

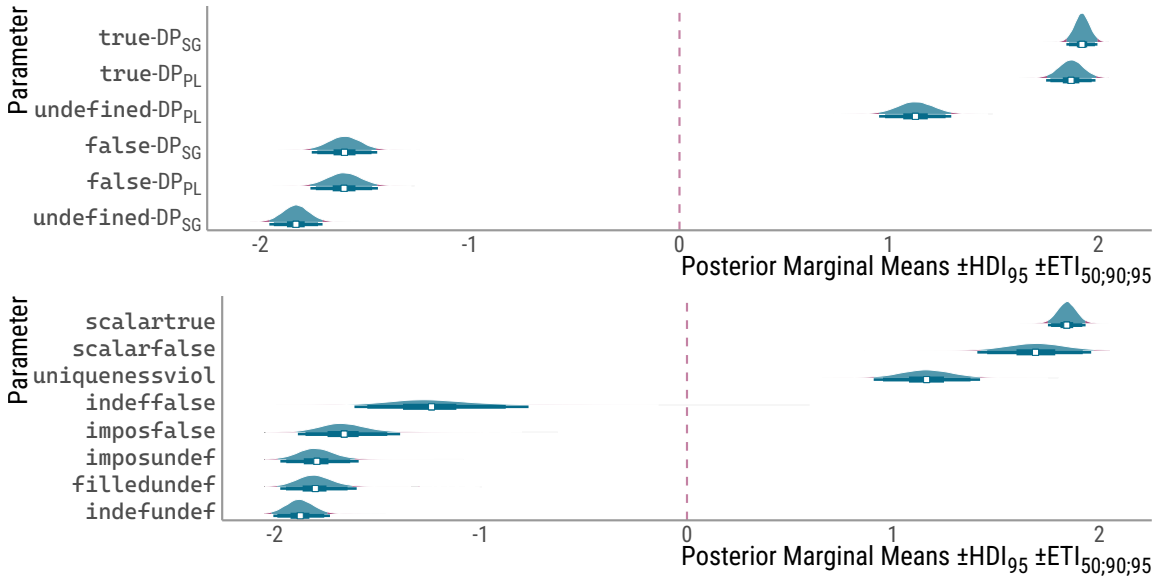


Figure 6: Location estimates for the critical (*top*) and control (*bottom*) conditions. Error bars show 50%, 90%, and 95% equal-tailed credible intervals (ETI), which indicate the central mass of the posterior with equal probability in each tail. The red areas of the distributions show the portions of the estimate that exceed the 95% highest-density credible interval (HDI), which include the most probable values of a parameter with highest density.

Turning towards the scale estimates now, which are displayed in Figure 7, we see a rather less clear-cut picture. For the controls, we do see that the `undefined` scenario for definite plurals has larger standard deviation estimates than the `undefined` scenario for definite singulars. However, `false` scenarios either show larger standard deviations (definite singular) or are at least on par (definite plural). The false controls, as well as the uniqueness violations, pattern similarly by displaying more dispersion.

5.6 Discussion

First, we may conclude that the method, at least in terms of the location estimates, worked: the results for the definite plurals across the three conditions lead to three distinct ratings, which mirror what we expected to see given the background in Križ & Chemla (2015). Hence, we may conclude that our results for these conditions represent a partial replication. The continuous version of the trivalent truth-value judgment task is thus similarly able to detect truth-value gaps, while at the same time recovering binary truth values.

As for the descriptions involving definite singular descriptions, we find that the `undefined` scenario did not pattern the way we expect it to on the assumption of a truth-value gap. Since the estimates patterned with those of false controls, we have no evidence in these data supporting a presuppositional analysis of existence. Uniqueness violations, on the other hand, lead to results that resembled the gappy judgments for homogeneity violations. On this interpretation, it seems that the current experiment conflicts with the predictions of both a pure Fregean and a Russellian approach to the semantics of singular definites. To the extent that the results we found are reliable, they seem to be broadly in line with Coppock & Beaver (2015), who predict that only uniqueness is lexically presupposed by the singular *the*. They also predict, however, that there should be a difference between argumental definites and predicative definites such that the former may have, depending on the choice of type shifter, an optional existence presupposition. Predicative definites, on the other hand, should never license one. The present experiment, however, did not include empty predicate definites and thus any difference between the two types of definites could not be tested.

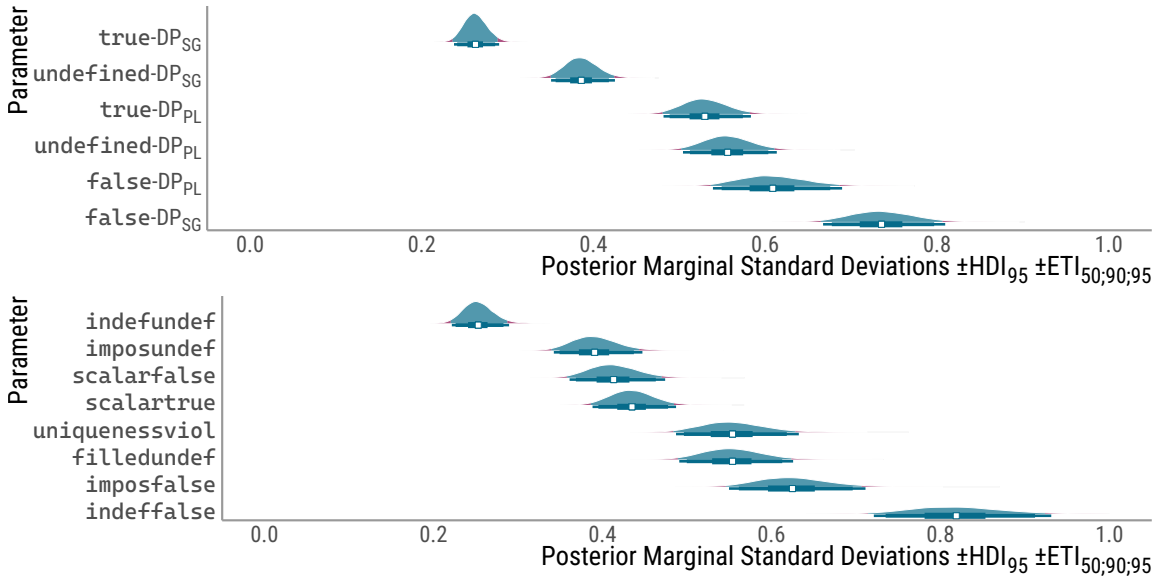


Figure 7: Back-transformed scale estimates for the critical (*top*) and control (*bottom*) conditions.

Regarding the Strawson (1964) and Schoubye (2009) view that topicality influences squeamishness, the fact that the syntactic manipulation had no effect on truth-value judgments suggests that no such dependency existed in the present experiment.

Originally, the experiment was supposed to test, among the predictions that the theoretical accounts make, also the potential for diagnosing presupposition violations via standard deviations on account of the triggered squeamishness. I called this the variance hypothesis. Unfortunately, the current version of the experiment found a hardly interpretable landscape of variance indices. Two reasons for this come to mind.

First, there were fewer of each control condition in (33) to (39) compared to the critical items. Since the quality of variance estimation is determined by the available data, this discrepancy calls into question whether participants saw enough of the controls to accurately estimate their standard deviation for these conditions. The sample standard deviation, at low sample sizes, is a biased estimator for the corresponding value in the population (contrary to the sample mean, which is unbiased). In order to obtain reliable estimates, the amount of data for each of the conditions should be the same.

Second, as was flagged throughout the presentation of the first experiment, the linguistic material includes the word *direkt* 'directly', which was especially highlighted in the introductory portion of the experiment. Notably, the instructions accompanying Figure 2 might have led participants towards a strategic judgment pattern rather than a linguistically determined one. Following the instructions, it seems plausible that participants may have resorted to only checking the cells that were directly adjacent to the mentioned shape in the relatum, which they plausibly would not have done if the instructions had not prompted such behavior. This kind of strategy would of course also make it hard to check the verification-based approaches in Lasersohn (1993) and von Fintel (2004), since biasing the way in which participants check the truth of an item could influence the applicability of secondary strategies.

All in all, while the results look promising at first blush, some of the design choices of this experiment prevent a direct interpretation of the results. In the second experiment, we will attempt to remedy these shortcomings. As the second experiment will be identical to the first one except for the modifications, we will only present the points of difference in detail.

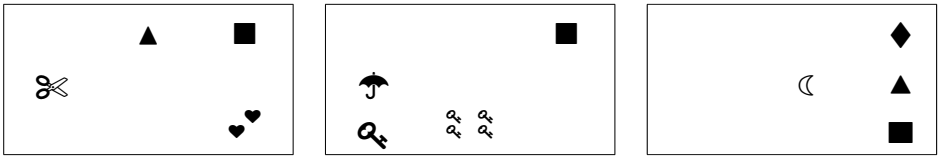
6 Experiment 2

6.1 Design and materials

In this experiment, we removed *direkt* from the item text and modified the images so that its inclusion was not necessary to begin with. We also decided to remove the lines of the matrix to further prevent strategies where participants only check cells adjacent to the relatum shape. The hope was that less strongly partitioned visual materials would discourage partial verification strategies and encourage checking the entirety of the images. The adapted linguistic and visual stimuli are presented below:

(43) a. Das Dreieck steht links neben dem Quadrat.

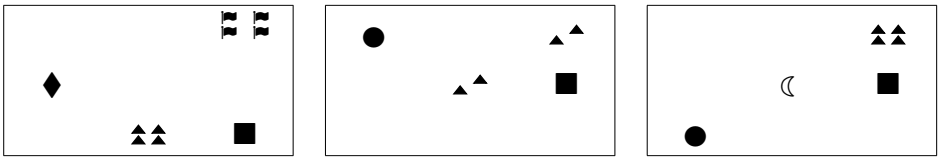
b. Links neben dem Quadrat steht das Dreieck.



`true` `undefined` `false`

c. Die Dreiecke stehen links neben dem Quadrat.

d. Links neben dem Quadrat stehen die Dreiecke.



`true` `undefined` `false`

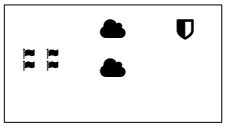
Because we decided to increase the number of items per condition, we decided to include fewer conditions to avoid making the experiment too long and hence too taxing for the participants. Participants saw each control condition six times per syntactic variant, totaling 84 items.

Since it is possible that the word order manipulation is not sufficient to bring out the differences predicted by Strawson (1964) and Schoubye (2009), we also included conditions where the relatum shape inside the PP (rather than the subject shape) was not shown in the data. This was done to compare with the singular `undefined` case from the critical items. Potentially, the difference in syntactic role—similar to the voice manipulation in (16)—is more appropriate for a comparison between topical and non-topical violations of existence than the word order manipulation alone, which had no effect in the first experiment. These new conditions are shown in (48) and (49). Lastly, we also included (50) among the fillers, which represents a uniqueness violation in a cumulative scenario, whose usefulness we will return to below. Below we list an example item (in both word orders) for each control condition together with the accompanying image.

(44) `uniquenessviol`

a. Die Wolke steht links neben dem Schild.
the cloud stands left next.to the sign

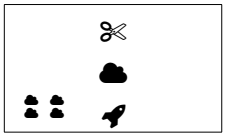
b. Links neben dem Schild steht die Wolke.
left next.to the sign stands the cloud



(45) `scalartrue`

a. Einige der Wolken stehen links neben der Rakete.
some of.the clouds stand left next.to the rocket

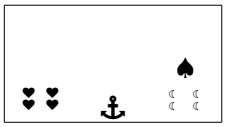
b. Links neben der Rakete stehen einige der Wolken.
left next.to the rocket stand some of.the clouds



(46) `scalarfalse`

a. Einige der Halbmonde stehen rechts neben dem Anker.
some of.the half.moons stand right next.to the anchor

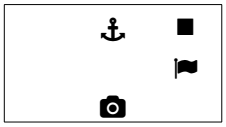
b. Rechts neben dem Anker stehen einige der Halbmonde.
right next.to the anchor stand some of.the half.moons



(47) `indefundef`

a. Eine Wolke steht links neben der Kamera.
a cloud stands left next.to the camera

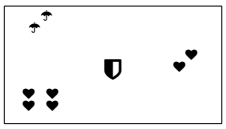
b. Links neben der Kamera steht eine Wolke.
left next.to the camera stands a cloud



(48) `relatumundef`

a. Der Schild steht über der Sonne.
the sign stands above the sun

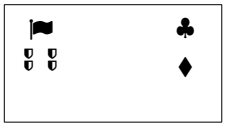
b. Über der Sonne steht der Schild.
above the sun stands the sign



(49) `relatumundefpl`

a. Die Schilde stehen links neben dem Anker.
the signs stand left next.to the anchor

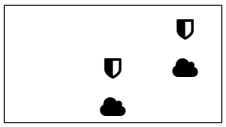
b. Links neben dem Anker stehen die Schilde.
left next.to the anchor stand the signs



(50) `doublepl`

a. Die Wolken stehen unter dem Schild.
the clouds stand under the sign

b. Unter dem Schild stehen die Wolken.
under the sign stand the clouds



6.2 Participants and procedure

We again tested 24 participants (mean age 22.5 ± 3.9, 21 female-identifying) under the same conditions as in the first experiment.

The procedure was altered in two ways: the potentially biasing instruction images were no longer included, and only the single warm-up item with a brief explanation was retained after a brief description of the experimental task and the scale. During the presentation of the experimental stimuli, the linguistic material was presented first, and participants had to actively press the space bar to make the image appear. This was done to give participants enough time to read the sentence without being distracted by checking for the symbols in the image. We also hoped that this would allow for a stronger focus on the word order manipulation. Apart from these changes, the procedure remained the same, as did the kind of data we collected throughout the experiment.

6.3 Results

Just like with the first experiment, the reaction times, the data for time to first motion of the slider, and the number of direction changes while using the slider are shown in the appendix in Figure 13.

Encouragingly, for the critical items shown in **??**, the model retrieved the means we saw in experiment 1, with the `true` and `false` conditions being rated as expected, and the `undefined` condition being rated as gappy with definite plurals; though with an overall more medial mean compared to the first experiment. As before, the `undefined` scenario for singular definites patterned with the `false` scenarios, and unlike the homogeneity violations. These results also follow from the posterior marginal means based on the location portion of the distributional model in **??**—see Section A.2 for complete model tables. Once again, the word order manipulation in the critical and the control items seems not to have had any reliable effect on the ratings.

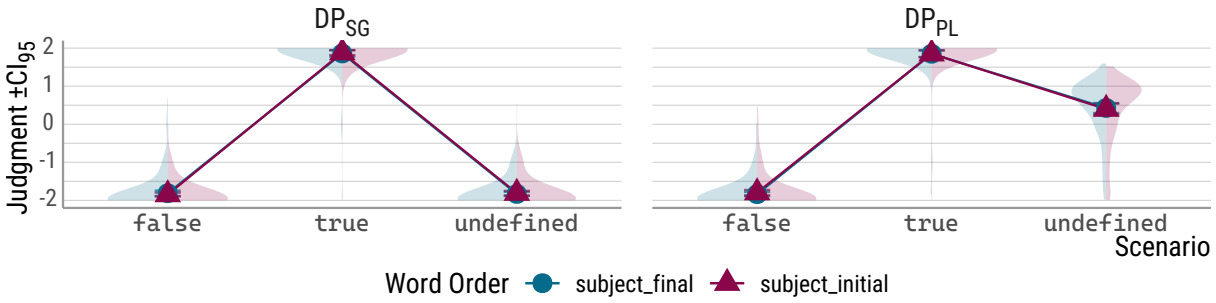


Figure 8: Overall ratings for the items in the critical portion of the second experiment.

As for the controls, we find means that let us suspect truth-value gaps for the `uniquenessviol`, the `scalarfalse`, and the `doublepl` condition. These fell in the medial region of the scale and mirrored the homogeneity violation condition. As an encouraging difference between the first and the second experiment, the `scalartrue` condition here received a rating that differed from the scalar implicature violations, suggesting that participants were sensitive to this distinction. The indefinite control for the existence violation condition, `indefundef`, received the expected falsity judgment and resembled the means we found for the `undefined` scenario with singular definites.

The two newly included controls where existence in the predicate was violated, `relatumundef` and `relatumundefpl`, were also rated as being strictly false, aligned with the `false` scenarios from the critical items and the `indefundef` control condition. As before, these impressions of the raw data are confirmed using the posterior marginal means from the models, Figure 10.

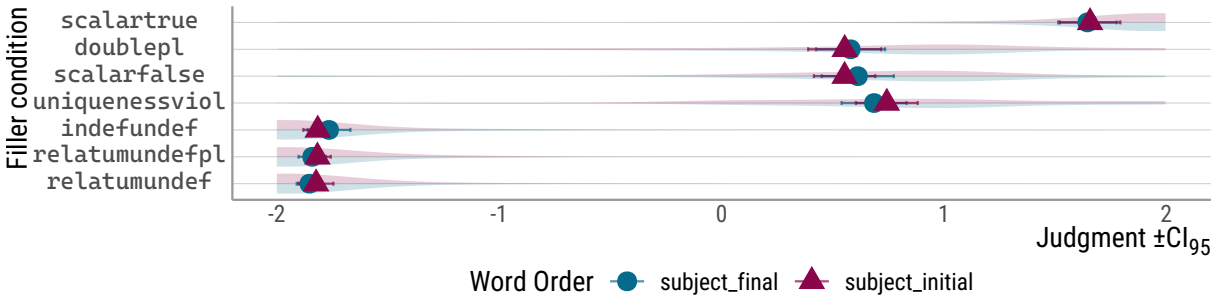


Figure 9: Overall ratings for the control items of the second experiment.

Let us now turn to the standard deviations. Just like with the first experiment, we will rely on posterior marginal effects for the comparisons and not attempt to draw any conclusions from the raw rating plots (indicated via the shaded areas) in Figures 8 and 9. Beginning with the upper panel of Figure 11, we can see that the homogeneity violations display the largest amount of uncertainty, together with the `true` plural definites, in the critical items. Thus, we find the predicted correspondence between scale-medial means and inflated variances. The condition that displayed the lowest standard deviations was the existence violation condition.

In the control items, we see that four conditions are roughly in line with the homogeneity violations from the critical portion of the experiment: the two conditions involving scalar inferences (or, at the very least, their triggers) and the two involving a uniqueness violation, `uniquenessviol` and `doublepl`. Here, the uniqueness violations and the violation of the scalar inference pattern are as expected from the perspective of the variance hypothesis. The `scalartrue` condition instead shows similar results to the `true` scenario with definite plurals featuring a mean that indicates a true rating and more spread-out rating behavior giving rise to that mean. We will return to a possible explanation for this correspondence below. The remaining three conditions in the controls, namely the two existence violations in predicative singular definites and the empty indefinite description, patterned like the `false` scenarios from the critical items in suggesting low uncertainty.

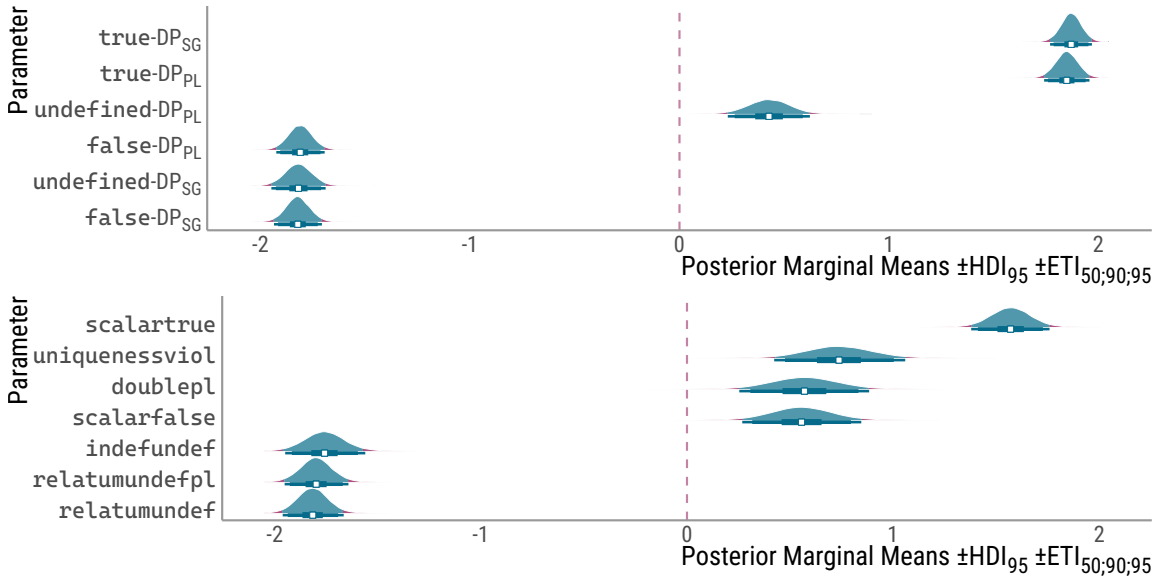


Figure 10: Location estimates for the critical and control conditions.

7 General Discussion

First of all, it is important to highlight that our minimum threshold for the validity of the experimental results for truth-value gaps was cleared: the gappy control, the `undefined` condition with definite plurals where homogeneity was violated, showed both relevant diagnostics for squeamishness: a scale-medial location estimate and a large standard deviation estimate. Our location results are not only a partial replication of Križ & Chemla (2015) but the (reliable) patterns we found are also stable across the two experiments. With the weaknesses of the first experiment in mind, we will base our discussion on matters relating to standard deviations, especially regarding the variance hypothesis and the question whether pragmatic processes are active, on the second experiment. The location-based results relating to the truth-value intuitions of the critical material are aligned between the two studies, and can thus be interpreted as mutually affirming. As for the controls in the first experiment, the estimated means there I consider less reliable as well based on the lack of precision in the estimates themselves; see the bottom panel of Figure 6.

7.1 The variance hypothesis

Since we will be using the standard deviation estimates from our models as the basis to argue for squeamishness and in turn as a diagnostic for intuitions of presupposition failure, we will start by discussing whether the variance hypothesis is substantiated by our experiments. Recall that we predicted that scale-medial judgments, which in this design stem from truth-value gaps, should come with increased uncertainty as measured by standard deviations. This prediction was borne out for the control portion of the first experiments and the entirety of the second experiment.

Before concluding that we found empirical support for the variance hypothesis, let me respond to potential lines of criticism. It is tempting to think that the larger variances we found with gappy conditions are a direct result of the way we set up the scale: a mean in the middle of the scale simply allows for a much larger amount of spread while still maintaining that same mean on a bounded scale. While this is certainly true in the general sense, in our experiment we have good reasons to think that the scale did not limit the amount of variation around means at the extremes. If that were the case, we should not have been able to find large standard deviation estimates with the `scalartrue` condition, which was rated as unmistakably completely true. The same holds for the `true` scenario with definite plurals, where we also find both a mean at the outer edge of the scale and a large amount of uncertainty. With this counterargument dealt with, we can conclude from the reliable data in the two experiments that the variance hypothesis holds. Participants who detect presupposition failure experience Strawsonian squeamishness, which manifests

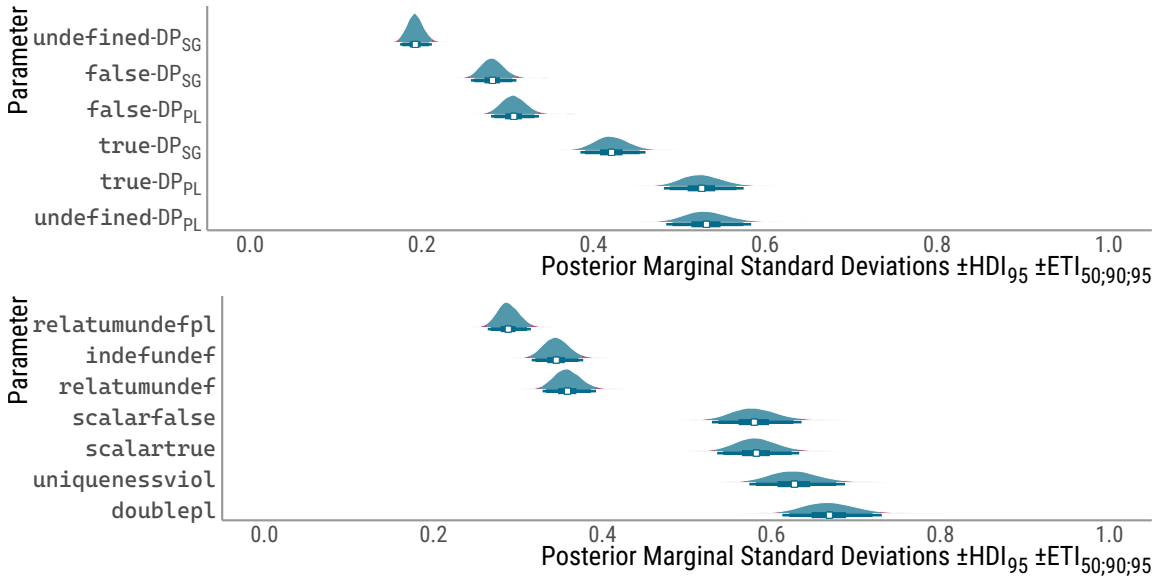


Figure 11: Back-transformed scale estimates for the critical and control conditions.

itself in their judgments by increasing the standard deviation in these conditions relative to non-gappy controls. In addition to the actual truth-value judgments we collected from participants over the course of these experiments, we can thus also make use of the (lack of) consistency in these ratings to diagnose intuitions of presupposition failure. This, as far as I am aware, has not been found before.

Two patterns that do not at first glance confirm the predictions: the scale-medial judgment we found for violations of scalar implicatures and the increased variances in the scenarios where homogeneity and scalar implicatures held. We will return to this after dealing with the major predictions for the presupposition status of existence and uniqueness.

7.2 The presupposition(s) of singular definites

Starting with existence, at first blush, our experimental results suggest that any analysis of singular definites according to which existence violations lead to a truth-value gap is not borne out by the data. The singular definite `undefined` condition did not pattern like the homogeneity violations, casting doubt on the existence of a truth-value gap in this case. The location estimate was on par with both the `false` scenarios in the critical items and the indefinite controls like `indefundef`, (47). In terms of dispersion, the estimated standard deviation was not only smaller than with the gappy controls, it was, together with the `indefundef` condition, the smallest of all conditions, indicating that participants were rather certain about their judgments for this condition across the experimental run. On the flip side, all conditions where uniqueness was violated showed double evidence of squeamishness: scale-medial overall judgments and inflated dispersion of the judgment distribution. Further, we found correspondence between uniqueness violations and the gappy baseline, the homogeneity violation condition.

Neither a Russellian semantics where neither existence nor uniqueness are presupposed nor a Fregean one where both are thus seems to be borne out by the data. Given our strongly asymmetrical results, an approach where existence and uniqueness come out as distinct meaning components is favored. A natural candidate is the account in Coppock & Beaver (2015) which in terms of lexical semantics predicts exactly the pattern we found.

There are, however, other experimental results that do not follow quite as naturally from the interplay of lexical semantics and type-shifting that Coppock & Beaver (2015) propose. Recall that argumental definites and predicative definites differed in their account because the former may optionally give rise to an existence presupposition via the route of the presuppositional type-shifter IOTA. Argumental definites that EX applies to and predicative definites, by contrast, are not predicted to license existence-related

truth-value gaps. In our experiment, all DPs where existence was violated were of argumental type. Now, Coppock & Beaver (2015) argued that type-shifting argumental definites via IOTA should be preferred over the option with EX because IOTA leads to the simpler type *e*, compared to the quantificational output that EX achieves. That is, contrary to what we found in our experiment where there was no indication that participants detected any difference between false controls and existence violations, we should have seen at least some amount of squeamishness triggered by the application of the preferred IOTA type-shifter.

However, Coppock & Beaver (2015) also discuss ways of departing from the default preference for IOTA and thus interpretations where existence is presupposed. For them, contextual features may prevent a presuppositional interpretation, forcing a parse with EX rather than IOTA. For our experiment, on the assumption that speakers prefer a judgment like 'completely false' over a gappy one and that they are willing to reparse an experimental item with EX when encountering an existence violation without any indication that this reparse occurred, our results are expected.

All in all, it seems that the analysis proposed by Coppock & Beaver (2015) is the most adequate for our data, yet it bears repeating that there was no indication in any of the conditions or for any of the metrics that existence was treated as presuppositional by the participants. That is, there was no support for the IOTA aspect of their proposal at all; a hypothetical variant of their proposal where IOTA is not available or does include any additional presuppositions would have been just compatible with our findings.

In the next steps, we will discuss whether the approaches that place the burden not on the lexical semantics of singular *the* but on information structure and verification strategies in the case of presupposition failure are corroborated by our results, and reject both as ill-supported.

7.3 Topicality

Since our data included no evidence for a presuppositional status of existence, there is automatically very little support for analysis according to which existence is only sometimes presuppositional, including the topicality and Question-Under-Discussion-based analyses in Strawson (1964) and Schoubye (2009), respectively. Neither the word order manipulation nor the existence violations in the `relatumundef` and `relatumundefpl` conditions in the second experiment showed any effect that allows for interpretations such that topic status or the (implicit) QUD had any effect on intuitions of presupposition failure. As we have seen, all instances of existence violations were judged as mere false sentences by the participants in the experiments, and contrasts like the one in (16) below could not be experimentally verified.

(16) a. The king of France visited the Exhibition yesterday. (#)
b. The Exhibition was visited by the king of France yesterday. (false)

While Strawson (1964) and Schoubye (2009) do not discuss uniqueness violations, on the assumption that their proposals extend to other kinds of presuppositions as well, we can also conclude that, even in cases where we do have evidence of presuppositionality, the word order variants were judged identically, contrary to what is expected from the extended proposals.

As for homogeneity, using example (iv) (Footnote 13), Križ (2015) argues that a contrast similar to what is predicted by Strawson (1964) and Schoubye (2009) for existence violations also holds for non-homogeneous scenarios with definite plurals. Here, too, we find no evidence of such an effect.

Hence, the hypothesis that truth-value gaps are only catastrophic (in terms of leading to squeamish 'neither nor' judgments) with topics or with certain QUD finds no support in our experiments.

7.4 Revisions

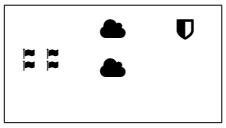
In the analysis of von Fintel (2004), presupposition failure does not necessarily lead to judgments of squeamishness or presupposition failure, but may be pragmatically shifted towards binary truth-value intuitions if the context allows for it. Given that our materials should in principle allow for such a process to apply, we have to ask whether our results are expected under the assumption that existence is presupposed but judgments correspond to pragmatic truth-value intuitions rather than the truth values that semantics yields.

One argument to doubt that a pragmatic process is the mechanism that drives the non-gappy judgments for existence is that it is hard to conceive of an in principle optional process that would not leave any traces in the data. Crucially, however, the existence violations and the `indefundef` conditions are not only parallel in terms of their location estimates, they also show very similar standard deviations. Hypothesizing that belief revision is needed to remove intuitions of presupposition failure with both the `undefined` and the `relatumundef(pl)` conditions, we expect at least increased variances, and potentially even effects on the location estimates. As we do not find either, I take it that the data do not support the idea that an existence presupposition is violated but intuitions are pragmatically adjusted as argued by von Fintel (2004).

But suppose now that the revisionist framework is correct, and that it applies to existence without any traces that the current experiments could detect. That is, let us assume that even though the kind of revision that von Fintel (2004) describes does not pattern with pragmatic processes and it does not have any effect on the ratings in the graded truth value design we employed. In this case, it should be possible to find the results we did, namely that existence violations look like they lead to mere falsity. This then invites the question whether our uniqueness violation data could also be explained under the assumption of the system in von Fintel (2004). What may seem counterintuitive given the clear indication of squeamishness in the data may appear less surprising if we consider the features of the uniqueness violation condition. In particular, it could be possible that belief revision in our experiment does not predict a pragmatic adjustment for uniqueness violations.

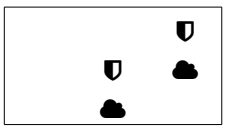
Recall that the revisionist analysis demands as a precondition for pragmatic adjustment that there is some verification foothold allowing for a binary evaluation of the utterance once the presupposition failure is removed from the information state. For this to hold without question, we expect for both clouds in the `uniquenessviol` in (44), repeated below, to clearly satisfy (or falsify) the predicate of being next to the shield. Here, only the top cloud shape verifies the predicate without question; the lower is arguably a less evident case. In other words, maybe the way we set up some of the visual materials prevented belief revision from adjusting away from squeamishness. In what is to come, I will argue that our experimental results conflict with the predictions von Fintel (2004) makes for violations of uniqueness presuppositions regardless.

(44) Die Wolke steht links neben dem Schild.
the cloud stands left next.to the sign



To see why, consider the `doublepl` control from (50), repeated below. The visual material features two shields, the definite description *dem Schild* is singular, however. At first sight, this is the configuration of uniqueness violations, but with a relatum that is interpreted cumulatively. Nevertheless, the resulting presupposition failure may lead to a pragmatic truth value following von Fintel (2004) if upon ignoring the presupposition violation a binary evaluation of the utterance is possible given the context. In the case below, on a cumulative interpretation of the plural referring to the clouds, we expect a TRUE judgment if we ignore the uniqueness violation because the shields are indeed under the clouds and satisfy the predicate. Hence, if the revisionist framework is correct, the `doublepl` condition should pattern with the `undefined` condition for singular definites, just at the opposite end of the scale, including a low standard deviation; see again Figures 10 and 11.

(50) Die Wolken stehen unter dem Schild.
the clouds stand under the sign



Yet, participants rated this condition as on par with the gappy homogeneity violations, with scale-medial judgments and large variance estimates, indicating that pragmatic truth value adjustments are not possible in this case. Even if we grant, contrary to what I have argued before, that the existence violation judgments are the result of pragmatically adjusted truth-value intuitions, the difference between the `undefined` condition for singular definites and the `doublepl` condition is striking. On the general interpretation of the system in von Fintel (2004) according to which it applies both to existence and uniqueness,

there is no reason to think that the two conditions should differ in their potential for feeding the pragmatic process described by von Fintel (2004). From this I conclude that von Fintel (2004) is unlikely to be operative for the judgments of falsity we found with violations of existence.

In sum, our experiments support an analysis of singular *the* that includes a uniqueness but not an existence presupposition. In addition, we found no evidence in support of secondary processes positing that only some instances of presupposition failure lead to squeamishness.

Before concluding, I will address two additional issues. One deals with a potential alternative look at the difference between existence and uniqueness violations by assuming that existence is soft and uniqueness a hard presupposition. The other concerns the until now unexplained similarity between the scalar implicature and homogeneity conditions in the second experiment. We will go through these in turn.

7.5 On gaps and other gaps

One alternative way of thinking about these results is in terms of a difference between truth-value gaps and presuppositions.[15] Given our results, it would be possible to assume that homogeneity violations and uniqueness violations are truth-value gaps (but not presuppositions), while existence violations are mere presuppositions. Supplemented with the assumption that our experiment design only elicits scale-medial judgments and squeamishness for truth-value gaps, this would explain the contrast between existence on the one hand and homogeneity and uniqueness on the other. Križ (2015) entertains the possibility that homogeneity differs from standard presuppositions in various ways which motivate the kind of distinction we're after. If it turns out that uniqueness patterns with homogeneity, maybe a more parsimonious explanation for our findings is that uniqueness is a truth-value gap, rather than a presupposition, as well. Plausibly, one could argue that what we call homogeneity with definite plurals is the same thing as uniqueness but in the singular. That is, while homogeneity is an all-or-nothing inference with plurals, a corresponding all-or-nothing inference with singulars is weak uniqueness.[16] If this is the case, the overlap between the two is unsurprising, while the independent existence inference is not necessarily expected to pattern accordingly.

The first diagnostic that espouses a difference between homogeneity and presuppositions that Križ (2015: p. 39) notes has to do with the hallmark feature of presuppositions, namely projection. As expected, the factive presupposition triggered by *know* appears to project from the antecedent of the condition in (51a); by contrast, Križ (2015: p. 39) denies that an all-or-nothing homogeneity inference projects in (51b). On a presuppositional treatment of homogeneity, a projective reading should be available, however.

(51) a. If Mary knows that John bought the ring, she's probably angry.
   ⇝ John bought the ring.
   b. If the subjects are asleep, the study can start.
   ⇝ Either all or none of the subjects are asleep.

The second such contrast that Križ (2015: p. 40) considers is shown in (52) where the pensive reaction indicated with *weell* is argued to be more appropriate in response to a homogeneity violation than to presupposition failure. Here, too, the expectation from a homogeneity-as-presupposition perspective is that both should align.

(52) a. *Context*: Adam has never smoked.
   A: Adam has stopped smoking.
   B: # Weeell…
   b. A: Adam has written the books.
   B: Weeell…

[15] I thank Clemens Mayr for pointing this out to me.

[16] This type of explanation runs into problems relating to the lack of parallelity between the singular and plural true controls in terms of standard deviations. We discuss this more in Section 7.7.

Leaving aside whether these differences are sufficient to assume a categorical difference between homogeneity and presuppositions and how this is to be cashed out theoretically, we will turn to uniqueness to see how it patterns. In line with the classical presuppositional view that Fregean approaches adopt, we see a clear pattern of projection from conditional antecedents, (53a). While my own intuitions are less robust for the response case in (54), the difference in the projection data is telling on its own:

(53) *Context*: There are many oak trees on campus.
   a. # If Mary recognized the oak tree on campus, she must've paid attention to the brochure.
   ⇝ There is (at most) one oak tree on campus.

(54) A: Mary saw the oak tree on campus.
   B: # Weeell…

Finally, even though this potential contrast between non-projecting truth-value gaps and projecting ones may serve to explain the contrasts in our experiments, it seems less clear why gaps without projection should be experimentally detectable while those that additionally project are not. Presumably, a meaning component that may lead to non-classical truth-values from embedded positions in addition to matrix positions should be easier to detect. This is also in line with the results in Thalmann & Matticchio (2024) who find that presupposition failures with *stop* and *again* in the complement of the attitude predicate *be certain* are detectable using the experimental method employed here, both with and without matrix negation. I conclude that pursuing an analysis according to which homogeneity, uniqueness, and existence are all truth-value gaps, but only one, namely existence, is a presupposition, is not promising in light of the empirical data.

7.6 Existence entailment or soft presupposition?

So far, I have presented the lack of an existence presupposition as the driver behind the contrast between the singular and the plural definite. While I will ultimately argue that this is the best explanation we have available, I now want to walk through some candidate explanations that identify other reasons for the absence of uniformity between existence and, say, uniqueness and homogeneity.

The first among these alternatives relies on the assumption that presuppositions are not consistent in terms of their projective properties. It is well known that there are contexts where there seems to be bifurcation of triggers such that one set of triggers, called hard, remains presuppositional while the other, called soft triggers, shows signs of what may be called presuppositional instability. [17] A hallmark feature of soft presuppositions is that they allow for non-projection in order to avoid the conversational breakdown that would follow from presupposition failure. To see what this means, consider the examples below from Abusch (2010):

(55) a. I have no idea whether John ended up participating in the Road Race yesterday. But if he won it, then he has more victories than anyone else in history.
   b. # I have no idea whether anyone read that letter. But if it is John who read it, let's ask him to be discreet about the content.

Both *win* and *it*-clefts are standardly understood to trigger presuppositions. In (55), the speaker asserts her ignorance of the presupposition and then embeds the triggering expression in the antecedent of a conditional. Given that presuppositions normally project out of conditional antecedents, we expect a contradiction with the ignorance assertion, but only find it for the *it*-cleft, not for *win*. The mechanism said to be responsible for avoiding the contradiction, presupposition suspension or local accommodation, is assumed to be possible for soft but not for hard presuppositions.

[17] I am trying to remain relatively non-committal here with respect to how this instability comes about. As far as I can see, multiple different options have been explored, from dedicated mechanisms that prevent a presupposition from projecting (suspension, local accommodation), to ones that do away with the presupposition in some way (sometimes called cancellation). Since I will argue that softness is not the way to go, commitment to a specific position seems unwarranted.

Against this background, one might now be tempted to conclude that if existence is a presupposition that allows for suspension (even if unembedded), this explains our results: a suspended presupposition should not be expected to trigger presupposition failure. Yet this interpretation relies on two assumptions, (i) that existence is a soft presupposition and (ii) that uniqueness, which did elicit presupposition failure, is a hard presupposition.[18] Past work has not settled on a definitive answer to either of these assumptions.

For Walker (2012), *the* is a soft trigger for both existence and uniqueness. For existence, he sides with Abusch (2002: p. 18) who notes in a footnote after discussing a case with an existence presupposition that the presupposition is suspendible, though without giving an example. Specifically for uniqueness, Walker (2012: p. 479) gives (56), arguing that without suspending the uniqueness presupposition, we should encounter infelicity. Judging (56) to be acceptable, he takes uniqueness to be a soft presupposition, together with existence.

(56) After the council, either the pope will unite Rome, or the popes will tear it apart.

Abbott (2006), on the other hand, gives the following judgment for a scenario with explicit ignorance, which seems to indicate that suspension is not available, favoring a hard trigger analysis.

(57) # Possibly no one owns this book, but if I find the owner, I will return it.

From this brief discussion of this topic in the literature, it seems like the suspension data above do not settle the matter. However, there exists another diagnostic that is sometimes used to differentiate between soft and hard presuppositions: projection from quantified environments. Charlow (2009), for example, argues on the basis of introspective data that one finds universal projection with strong triggers like *too* across the board, (58b) and (58d). With soft triggers like *quit*, he says projection patterns are sometimes non-universal, particularly when the embedding quantifier is not universal/negative existential, (58c).

(58) a. Each/none of these 100 students quit smoking.
b. Each/none of these 100 students smokes Marlboros too.
c. Some of these 100 students quit smoking.
d. Some of these 100 students smoke Marlboros too.

Despite these claims, recent experimental results by Thalmann & Matticchio (2024) favor a more cautious application of this alleged diagnostic. For projection from the attitude predicate *(not) be certain*, they find no difference between the soft trigger *stop* and the hard trigger *again* in terms of projection, no matter whether matrix negation was present or not. Instead, they find uniform projection; a pattern that aligns with the predictions of Strong Kleene logic as discussed in Fox (2013) (who builds on George 2010), though without the necessity for assuming additional projection tampering mechanisms like local accommodation. These results substantially weaken the case for this diagnostic, and we will leave introspective tests of projective strength from quantifiers aside for that reason.

Finally, there have been experimental results pointing towards a difference between soft and hard triggers in terms of at-issueness. Chen, Thalmann & Antomo (2022) test the acceptability of responding to a question with answers where only the presupposed meaning components provide the requested information. Focusing on their results for adults, they find that hard triggers are penalized substantially in this paradigm, patterning together with non-restrictive relative clauses, (59). The presuppositions of soft triggers, however, were judged as less marked than hard presuppositions but more marked compared to control conditions where the assertion provided the answer. The authors conclude that the ability to address questions can serve as a genuine diagnostic between soft and hard presuppositions.

(59) What did the duck get? — # The duck, who got some cough syrup by the way, is in bed now.

[18] There is additional complexity that lurks here: suspension/local accommodation would have to be able to target only one of the presuppositions triggered by singular the and leave the other one unaffected. On the assumption of an operator-based analysis for local accommodation (as in Strong Kleene logic, see Beaver and Krahmer 2001), it is not trivial to achieve the necessary selectivity.

It has been argued that at-issueness is a gradient phenomenon by Tonhauser, Beaver & Degen (2018). To the extent that this is correct,[19] it impacts the usefulness of both the diagnostic in Chen, Thalmann & Antomo (2022) and the relevance of non-experimental judgments for deciding between a soft or hard classification on the basis of at-issueness. Since Chen, Thalmann & Antomo (2022) did not include the among their list of presupposition triggers, it is hard to decide whether (versions of) (60) show the same acceptability contrast as (61), which was included in the experiment.

(60) a. How many popes are there? — The pope resides in the Vatican.
b. Is there a pope? — The pope resides in the Vatican.

(61) a. Did the duck participate in the competition? — The duck won the competition.
b. # Did the panda score a goal for the first time yesterday? — The panda scored a goal again yesterday.

Another point to doubt that the contrast we obtained between existence and uniqueness violations is due to a qualitative difference between the presuppositions lies with the second metric we used to identify presupposition failure: increases in standard deviations. Recall that violations of existence did not lead to any increase in variance; in fact, the undefined scenario with singular DPs was among the conditions with the lowest variance in the entire experiment, exceeded not only by other presuppositional conditions but also non-presuppositional controls. If we take Strawson (1964) seriously and commit to the view that experimental participants experience squeamishness when confronted with presupposition failure and that increased variance indices are one way squeamishness is measured, then the results for existence force us to accept either of two conclusions: (i) existence is not presupposed or, on the assumption that existence is a soft presupposition, (ii) the violation of soft presuppositions does not lead to detectable increases in variance. Thalmann & Matticchio (2024), in their study of the projection problem with attitudes, used the same experimental method we did, and found no difference between their hard and their soft triggers in terms of variances, with both violations resulting in larger variances than for conditions without presupposition failure. This suggests that (ii) is not borne out empirically.

Finally, von Fintel (2004) also argues against the involvement of local accommodation based on the outcome of the 'Hey, wait a minute' diagnostic. Recall (25), repeated below, where the local accommodation predicts that the 'Hey, wait a minute' should lead to infelicity, contrary to fact:

(25) A′: The Exhibition was visited by the king of France yesterday. (false)
B′: Hey, wait a minute—I had no idea France was still a monarchy.

In sum, interpreting the difference between uniqueness and existence conditions as stemming from presupposition hardness is not without its complications, and I have tried to show that both our results and the results in other experiments speak against this interpretive avenue. These challenges are compounded by the ongoing controversy surrounding the classification of these two inferences (granting their presuppositional status) as soft or hard. Given the rejection of softness as a viable explanation for our data, I instead conclude that existence violations did not pattern like one would expect presupposition failures to pattern.

7.7 Variance and definite plurals

Before concluding, I want to discuss a finding which may be of interest despite being unrelated to the express goals of this study. In the second experiment, we found larger standard deviations with plural

[19] Gradience between items shows up frequently in experiments across all linguistic domains. The question is, however, whether the source of this gradience actually originates from the critical manipulation (i.e., the presuppositions of the different triggers) or whether it is caused by factors that are almost impossible to control for. Lexical differences, for example, very often lead participants to favor certain items over others, for no theoretically substantial reason at all. It is very hard to show that the observed item differences are directly caused by the characteristics of the presupposition rather than by the other denotational or even extralinguistic features of the triggering lexical element; let alone the other lexical material in the items.

definites both in `undefined` and `true` scenarios. At first, more uncertainty in the true controls is unexpected, since the utterance did not fall into a truth-value gap in this condition, and so the variance hypothesis does not predict squeamishness-induced variance hikes.

However, maybe this could be interpreted as a sign of the involvement of a pragmatic process to derive the homogeneity inference in the first place. Taking this view, we could see the application of the pragmatic process, not the violation of an inference, as the driver behind the larger variance. In other words, this pattern would not be an instance of squeamishness but would merely reflect the application of a process which affects variance for independent reasons. As an example of one such account, consider the exhaust-based analyses of homogeneity in Magri (2014), Bar-Lev & Fox (2020), and Bar-Lev (2021). Broadly speaking, these proposals assume a weak, existential semantics of definite plurals which may optionally be strengthened via an implicature to derive the universal interpretation, which is violated in our `undefined` scenario. On the plausible assumption that deriving an optional implicature has an impact on standard deviations, our experimental results for the homogeneity items become less surprising even in the *true* scenario, where arguably such an implicature still derived, just not violated. This explanation receives further support from the correspondence with the `scalartrue` condition in the controls, which also showed large standard deviations.

Despite the appeal of such a view, it raises a variety of issues when adopted to explain our data. The first of these has to do with the obligatoriness of these inferences. In a system where both homogeneity and scalar implicatures are derived using a covert operator like in the exhaustification literature, one may wonder why speakers do not simply reparse the LF without that operator when the inference in question is violated. If the operator is simply a syntactic realization of Gricean reasoning, which is known to yield weak, cancellable enrichments, we may expect cancellation to apply in scenarios where otherwise a false sentence would arise. As Magri (2009) argues with examples like the one below, however, cancellation of scalar implicatures is not always possible. Even though (62) feels odd because of an implicature violation, an felicitous reading where the offending implicature that not all Italians come from a warm country is cancelled is unavailable. Because of this and similar data points, Magri and others assume obligatory exhaustification.[20]

(62) # Some Italians come from a warm country.

This obligatoriness assumption, however, conflicts with the standard deviation increases we are trying to explain: if the inclusion of the exhaust operator requires a reparse of the utterance with the operator, and if that reparse is what leads to increased variance, then inflated variances remain unexplained with obligatory exhaustification. So while we may be able to take the oddity we see in (62) as connected to inflated variances with implicature violations, no such violation was present in the `scalartrue` condition or the true controls with definite plurals.

Another way is to posit that the mechanism that underlies exhaustification—the exclusion and sometimes the inclusion of alternatives—affects variance in certain cases. Perhaps establishing/pruning (Bar-Lev 2021) the set of alternatives or the process of negating innocently excludable alternatives requires more cognitive effort, which leads to appreciable increases in standard deviations. If this is on the right track, standard deviations are expected to rise even when no violation occurs. The presence of the obligatory operator would be sufficient. This is a plausible explanation but one that requires independent support, and until such support is found, the above remains speculative at best and we are left with an unexplained pattern.

However, maybe the exhaustification analysis at least serves to explain what happens in the violation conditions, namely the scale-medial judgments and the exacerbated variance scores. At first glance, on a view where homogeneity and scalar inferences are derived pragmatically via a mere strengthening of the truth conditions, the overall correspondence with uniqueness violations is surprising: violating an implicature that derives a universal interpretation for definite plurals should lead to judgments of falsity rather than indications of a truth-value gap. The same is true for violations of the upper-bounded inference in

[20] This obligatoriness assumption extends also to the presuppositional exhaust proposals in Bassi, Del Pinal & Sauerland (2021) and Guerrini & Wehbe (2025) we will talk about below.

the `scalarfalse` condition, which is expected to trigger falsity rather than scale-medial judgments and squeamishness. So while exhaust-based accounts potentially have a handle on the larger standard deviations for homogeneity and scalar implicature, the question why violations result in scale-medial judgments remains unanswered on this view.

One option to make sense of these presupposition-like results is to adopt a view according to which scalar inferences (and exhaustification processes more generally) are not operative on the level of assertion but instead amount to presuppositions; an idea that is pursued in Bassi, Del Pinal & Sauerland (2021) and Del Pinal, Bassi & Sauerland (2024). One way to interpret our results for homogeneity inferences and scalar implicatures alike then is to assume that both are derived using the presuppositional implicature system in Bassi, Del Pinal & Sauerland (2021) and Del Pinal, Bassi & Sauerland (2024). Bassi, Del Pinal & Sauerland (2021) argue that this move from an assertive to a presuppositional mechanism is necessary for scalar implicatures, such that the upper-bounded implicature generates a truth-value gap rather than a more informative assertion. Guerrini & Wehbe (2025) argue that presuppositional exhaust is likewise responsible for deriving homogeneity inferences. Accepting both explains the scale-medial results and, via the variance hypothesis, the inflated uncertainty in the violation conditions for definite plurals and the scalar implicature items. Despite the success of a presuppositional view in explaining our homogeneity and scalar implicature data in the violation conditions, in conditions where no presupposition is violated, the assumption of a presuppositional mechanism should not lead to a prediction of squeamishness. Even after a move to presuppositional exhaustification, we still do not have a clear explanation for the standard deviations in the non-violation conditions. We are hopeful that future work into these questions will provide more definitive answers on the link between standard deviations and exhaustification-derived meanings.

8 Conclusion

To conclude our investigation, we have seen that there are good reasons to suspect that two inferences commonly associated with singular definites, existence, and uniqueness, are to be classified differently. In particular, the truth-value intuitions in the continuous trivalent truth-value judgment task for violations of uniqueness patterned together homogeneity violations while differing from classically true and false controls. Existence violations, on the other hand, were treated in parallel with false controls. These results are in conflict with both a Russellian, non-presuppositional semantics of singular definites as in Russell (1905) and a more standard Fregean view according to which both uniqueness and existence are presupposed (e.g., Heim & Kratzer 1998, Elbourne 2013). While we interpreted our results as being compatible with the asymmetrical account in Coppock & Beaver (2015), on their account, it remains an open question why there was no sign of presupposition failure for existence violations in our experimental data at all. We also considered whether a difference in presuppositional strength could be at work such that existence presuppositions are soft presuppositions and uniqueness presuppositions are hard. An alternative hypothesis could be that violating soft presuppositions leads to markedly different results in this experimental paradigm than the violation of hard presuppositions. For this classification, however, Thalmann & Matticchio (2024) found counterevidence in their experiment on presupposition projection from attitudes using the very same method, which included uncontroversial soft and hard triggers.

Second, we found no support for approaches to presuppositions according to which violations may be judged as classically truth or falsity when certain contextual or information structural conditions apply. Neither manipulations of word order nor of syntactic role had any effect on truth-value judgments or their variance in our experiments. This conflicts with both Strawson (1964) and Schoubye (2009). In addition, our data did align with the predictions of verification-based approaches like the one in Lasersohn (1993) and von Fintel (2004).

Lastly, our experiments support what I called the variance hypothesis, namely the assumption that presupposition failure should lead to a feeling of Strawsonian squeamishness in experimental participants, which is detectable via inflated standard deviations. Here, too, violations of uniqueness and homogeneity patterned together and contrasted with true and false controls as well as existence violations. Despite the

focus in experimental work on estimating measures of central tendency such as means and disregarding measures of dispersion like standard deviations as mere noise, the variance hypothesis allowed us to interpret that noise as a signal for the detection of presupposition failure. In turn, I hope this could become a further diagnostic in the experimental study of truth-value gaps, though this is reliant upon the effect on standard deviations to be present in other experimental paradigms as well. We leave the study of, say, acceptability judgment tasks from this perspective for future work.

References

Abbott, Barbara. 2006. Where have some of the presuppositions gone? In Betty J. Birner & Gregory Ward (eds.), *Drawing the boundaries of meaning: Neo-Gricean studies in pragmatics and semantics in honor of Laurence R. Horn*, 1–20. Amsterdam: Benjamins. https://doi.org/10.1075/slcs.80.02abb.

Abrusán, Márta. 2016. Presupposition cancellation: explaining the 'soft–hard' trigger distinction. *Natural Language Semantics* 24(2). 165–202. https://doi.org/10.1007/s11050-016-9122-7.

Abrusán, Márta & Kriszta Szendrői. 2013. Experimenting with the king of France: Topics, verifiability and definite descriptions. *Semantics & Pragmatics* 6. 10. https://doi.org/10.3765/sp.6.10.

Abusch, Dorit. 2002. Lexical alternatives as a source of pragmatic presuppositions. In Brendan Jackson (ed.), *Proceedings of Semantics and Linguistic Theory (SALT) 12*, 1–19. https://doi.org/10.3765/salt.v12i0.2867.

Abusch, Dorit. 2010. Presupposition triggering from alternatives. *Journal of Semantics* 27(1). 37–80. https://doi.org/10.1093/jos/ffp009.

Bar-Lev, Moshe E. 2021. An implicature account of homogeneity and non-maximality. *Linguistics & Philosophy* 44(5). 1045–1097. https://doi.org/10.1007/s10988-020-09308-5.

Bar-Lev, Moshe E. & Danny Fox. 2020. Free choice, simplification, and innocent inclusion. *Natural Language Semantics* 28(3). 175–223. https://doi.org/10.1007/s11050-020-09162-y.

Bassi, Itai, Guillermo Del Pinal & Uli Sauerland. 2021. Presuppositional exhaustification. *Semantics & Pragmatics* 14(11). https://doi.org/10.3765/sp.14.11.

Beaver, David I. & Emiel Krahmer. 2001. A partial account of presupposition projection. *Journal of Logic, Language and Information* 10(2). 147–182. https://doi.org/10.1023/a:1008371413822.

Bürkner, Paul-Christian. 2021. Bayesian item response modeling in R with brms and stan. *Journal of Statistical Software* 100(5). 1–54. https://doi.org/10.18637/jss.v100.i05.

Charlow, Simon. 2009. "Strong" predicative presuppositional objects. In *Proceedings of ESSLLI 2009*.

Chen, Yuqiu, Maik Thalmann & Mailin Antomo. 2022. Presupposition triggers and (not-)at-issueness. Insights from language acquisition into the soft-hard distinction. *Journal of Pragmatics* 199. 21–46. https://doi.org/10.1016/j.pragma.2022.06.014.

Coppock, Elizabeth & David I. Beaver. 2015. Definiteness and determinacy. *Linguistics & Philosophy* 38(5). 377–435. https://doi.org/10.1007/s10988-015-9178-8.

Coppock, Elizabeth & Lucas Champollion. 2018. *Formal semantics boot camp*.

Del Pinal, Guillermo, Itai Bassi & Uli Sauerland. 2024. Free choice and presuppositional exhaustification. *Semantics and Pragmatics* 17(3). https://doi.org/10.3765/sp.17.3.

Elbourne, Paul. 2013. *Definite descriptions*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199660193.001.0001.

von Fintel, Kai. 2004. Would you believe it? The king of France is back! (presuppositions and truth-value intuitions). In Marga Reimer & Anne Bezuidenhout (eds.), *Descriptions and beyond*. Oxford: Oxford University Press.

von Fintel, Kai. 2008. What is presupposition accommodation, again? *Philosophical Perspectives* 22(1). 137–170. https://doi.org/10.1111/j.1520-8583.2008.00144.x.

Fox, Danny. 2013. Presupposition projection from quantificational sentences: Trivalence, local accommodation, and presupposition strengthening. In Ivano Caponigro & Carlo Cecchetto (eds.), *From grammar to meaning*, 201–232. Cambridge University Press. https://doi.org/10.1017/cbo9781139519328.011.

Frege, Gottlob. 1997. Über Sinn und Bedeutung (1892). In Michael Beaney (ed.), *The Frege reader*, 151–171. Malden, MA: Blackwell Publishers.

George, B. 2010. A new case for an old logic: Reviving Strong Kleene approaches to presupposition projection. Unpublished ms. UCLA.

Geurts, Bart. 2008. Existential import. In Ileana Comorovski & Klaus von Heusinger (eds.), *Existence: Semantics and syntax*, 253–271. Dordrecht: Springer. https://doi.org/10.1007/978-1-4020-6197-4_9.

Guerrini, Janek & Jad Wehbe. 2025. Homogeneity as presuppositional exhaustification. *Journal of Semantics*.

Hajičová, Eva. 1984. Presupposition and allegation revisited. *Journal of Pragmatics* 8(2). 155–167. https://doi.org/10.1016/0378-2166(84)90046-8.

Heim, Irene. 1982. *The semantics of definite and indefinite noun phrases*. UMass Amherst dissertation.

Heim, Irene & Angelika Kratzer. 1998. *Semantics in generative grammar*. Oxford: Blackwell.

Horn, Laurence R. 1972. *On the semantic properties of logical operators in English*. Yale University dissertation.

Križ, Manuel. 2015. *Aspects of homogeneity in the semantics of natural language*. University of Vienna dissertation.

Križ, Manuel & Emmanuel Chemla. 2015. Two methods to find truth-value gaps and their application to the projection problem of homogeneity. *Natural Language Semantics* 23(3). 205–248. https://doi.org/10.1007/s11050-015-9114-z.

Lasersohn, Peter. 1993. Existence presuppositions and background knowledge. *Journal of Semantics* 10(2). 113–122. https://doi.org/10.1093/jos/10.2.113.

Lenth, Russell V. 2024. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.10.0.

Magri, Giorgio. 2009. A theory of individual-level predicates based on blind mandatory scalar implicatures. *Natural Language Semantics* 17(3). 245–297. https://doi.org/10.1007/s11050-009-9042-x.

Magri, Giorgio. 2014. An account for the homogeneity effect triggered by plural definites and conjunction based on double strengthening. In Salvatore Pistoia-Reda (ed.), *Pragmatics, semantics and the case of scalar implicatures*, 99–145. London: Palgrave Macmillan UK. https://doi.org/10.1057/9781137333285_5.

Partee, Barbara H. 1986. Noun phrase interpretation and type-shifting principles. In Jeroen Groenendijk, Dick de Jongh & Martin J. Stokhof (eds.), *Studies in Discourse Representation Theory and the theory of generalized quantifiers*, 115–143. Dordrecht: Foris. https://doi.org/10.1515/9783112420027-006.

Partee, Barbara H. 1996. Allegation and local accommodation. In Barbara H. Partee & Petr Sgall (eds.), *Discourse and meaning: papers in honor of Eva Hajičová*, 65–86. Amsterdam: John Benjamins. https://doi.org/10.1075/z.78.12par.

Peirce, Jonathan, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman & Jonas Kristoffer Lindeløv. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods* 51(1). 195–203. https://doi.org/10.3758/s13428-018-01193-y.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Russell, Bertrand. 1905. On denoting. *Mind* 14(56). 479–493. https://doi.org/10.5040/9781474216166.0007.

Schoubye, Anders J. 2009. Descriptions, truth value intuitions, and questions. *Linguistics & Philosophy* 32(6). 583–617. https://doi.org/10.1007/s10988-010-9069-y.

Strawson, P. F. 1950. On referring. *Mind* 59(235). 320–344. http://www.jstor.org/stable/2251176.

Strawson, P. F. 1964. Identifying reference and truth-values. *Theoria* 30(2). 96–118. https://doi.org/10.1111/j.1755-2567.1964.tb00404.x.

Sun, Siqi, Karen M. Schmidt & Teague R. Henry. 2025. *Don't Let Your Likert Scales Grow Up To Be Visual Analog Scales: Understanding the Relationship Between Number of Response Categories and Measurement Error*. Ms. https://doi.org/10.48550/ARXIV.2502.02846.

Parameter	ETI Median	$ETI_{2.5}$	$ETI_{97.5}$	HDI Mode	$HDI_{2.5}$	$HDI_{97.5}$
Intercept	-0.02	-0.09	0.05	-0.01	-0.09	0.05
SD-Intercept	-0.72	-0.77	-0.68	-0.72	-0.77	-0.68
false	-1.58	-1.67	-1.49	-1.58	-1.67	-1.49
undefined	-0.33	-0.4	-0.27	-0.34	-0.4	-0.27
DPSG	-0.48	-0.54	-0.42	-0.48	-0.54	-0.42
subject-final	-0.01	-0.04	0.01	-0.02	-0.04	0.01
false:DPSG	0.49	0.42	0.55	0.49	0.42	0.55
undefined:DPSG	-1	-1.07	-0.92	-0.99	-1.07	-0.92
false:subject-final	-0.01	-0.06	0.03	-0.01	-0.06	0.03
undefined:subject-final	0.02	-0.02	0.06	0.02	-0.02	0.06
DPSG:subject-final	-0.02	-0.05	0.01	-0.02	-0.05	0.01
false:DPSG:subject-final	-0.01	-0.06	0.03	-0.01	-0.06	0.03
undefined:DPSG:subject-final	0.01	-0.03	0.05	0.01	-0.03	0.04
SD-false	0.32	0.25	0.38	0.32	0.25	0.38
SD-undefined	-0.05	-0.11	0.01	-0.05	-0.11	0.01
SD-DPSG	-0.15	-0.19	-0.11	-0.14	-0.19	-0.11
SD-subject-final	0.08	0.04	0.12	0.08	0.04	0.12
SD-false:DPSG	0.24	0.18	0.3	0.24	0.18	0.3
SD-undefined:DPSG	-0.04	-0.09	0.02	-0.04	-0.09	0.02
SD-false:subject-final	-0.14	-0.2	-0.09	-0.15	-0.2	-0.09
SD-undefined:subject-final	-0.1	-0.16	-0.04	-0.1	-0.16	-0.05
SD-DPSG:subject-final	0.08	0.04	0.11	0.07	0.04	0.11
SD-false:DPSG:subject-final	-0.1	-0.15	-0.04	-0.09	-0.15	-0.04
SD-undefined:DPSG:subject-final	-0.05	-0.1	0.01	-0.05	-0.1	0.01

Table 1: Model summary for the critical items in Experiment 1. All $\hat{R} = 1$ and all effective sample sizes for bulk and tail $> 10k$. Standard deviation parameters are prefixed with 'SD-' and are reported on the log scale.

Thalmann, Maik & Andrea Matticchio. 2024. On being certain that presuppositions don't project universally. In *Proceedings of the Amsterdam Colloquium*, 378–385.

Thalmann, Maik & Andrea Matticchio. (to appear). How to be mistaken and still happy: Belief-Relative presuppositions and factivity illusions. In *Proceedings of the 48th penn linguistics conference*.

Tonhauser, Judith, David I. Beaver & Judith Degen. 2018. How projective is projective content? Gradience in projectivity and at-issueness. *Journal of Semantics* 35(3). 495–542. https://doi.org/10.1093/jos/ffy007.

Walker, Andreas. 2012. Focus, uniqueness and soft presupposition triggers. In Maria Aloni, Floris Roelofsen, Galit W. Sassoon, Katrin Schulz & Matthijs Westera (eds.), *Logic, language and meaning. 18th amsterdam colloquium*, 475–484. Berlin: Springer. https://doi.org/10.1007/978-3-642-31482-7_47.

Parameter	ETI Median	$ETI_{2.5}$	$ETI_{97.5}$	HDI Mode	$HDI_{2.5}$	$HDI_{97.5}$
Intercept	1.89	1.8	1.98	1.9	1.8	1.98
SD-Intercept	-1.1	-1.25	-0.95	-1.1	-1.25	-0.95
filledundef	-3.75	-3.92	-3.54	-3.75	-3.94	-3.56
imposfalse	-3.62	-3.84	-3.34	-3.61	-3.85	-3.36
imposundef	-3.71	-3.89	-3.49	-3.71	-3.9	-3.51
indeffalse	-3.22	-3.57	-2.74	-3.25	-3.61	-2.79
indefundef	-3.77	-3.92	-3.6	-3.77	-3.92	-3.6
scalarfalse	-0.16	-0.45	0.13	-0.16	-0.45	0.13
uniquenessviol	-0.65	-0.9	-0.39	-0.65	-0.9	-0.39
subject-initial	-0.1	-0.26	0.06	-0.1	-0.26	0.06
filledundef:subject-initial	0.2	-0.1	0.49	0.2	-0.1	0.49
imposfalse:subject-initial	0.21	-0.13	0.55	0.22	-0.14	0.55
imposundef:subject-initial	0.13	-0.12	0.38	0.12	-0.12	0.38
indeffalse:subject-initial	0.25	-0.15	0.65	0.24	-0.14	0.66
indefundef:subject-initial	0.09	-0.12	0.29	0.09	-0.11	0.29
scalarfalse:subject-initial	0.03	-0.23	0.3	0.02	-0.23	0.3
uniquenessviol:subject-initial	-0.07	-0.37	0.23	-0.08	-0.36	0.24
SD-filledundef	0.34	0.15	0.54	0.34	0.15	0.54
SD-imposfalse	0.48	0.27	0.68	0.48	0.28	0.69
SD-imposundef	0.08	-0.12	0.28	0.08	-0.13	0.28
SD-indeffalse	0.67	0.47	0.87	0.68	0.47	0.87
SD-indefundef	-0.25	-0.45	-0.05	-0.24	-0.45	-0.05
SD-scalarfalse	-0.03	-0.24	0.18	-0.03	-0.25	0.18
SD-uniquenessviol	0.18	-0.04	0.39	0.18	-0.04	0.39
SD-subject-initial	0.53	0.38	0.68	0.54	0.38	0.69
SD-filledundef:subject-initial	-0.21	-0.45	0.04	-0.22	-0.46	0.03
SD-imposfalse:subject-initial	-0.23	-0.48	0.02	-0.22	-0.49	0.01
SD-imposundef:subject-initial	-0.37	-0.63	-0.11	-0.36	-0.63	-0.11
SD-indeffalse:subject-initial	-0.07	-0.32	0.18	-0.05	-0.32	0.19
SD-indefundef:subject-initial	-0.59	-0.83	-0.33	-0.59	-0.83	-0.33
SD-scalarfalse:subject-initial	-0.04	-0.3	0.23	-0.03	-0.3	0.23
SD-uniquenessviol:subject-initial	0.13	-0.13	0.39	0.13	-0.13	0.38

Table 2: Model summary for the control items in Experiment 1. All $\hat{R} = 1$ and all effective sample sizes for bulk and tail $> 10k$. Standard deviation parameters are prefixed with 'SD-' and are reported on the log scale.

Parameter	ETI Median	$ETI_{2.5}$	$ETI_{97.5}$	HDI Mode	$HDI_{2.5}$	$HDI_{97.5}$
Intercept	-0.22	-0.28	-0.15	-0.22	-0.28	-0.15
SD-Intercept	-1.04	-1.08	-1	-1.04	-1.08	-1
false	-1.6	-1.68	-1.52	-1.6	-1.68	-1.52
undefined	-0.48	-0.57	-0.39	-0.48	-0.57	-0.39
DP-SG	-0.37	-0.41	-0.33	-0.37	-0.41	-0.33
subject-final	-0	-0.02	0.02	-0	-0.02	0.02
false:DP-SG	0.37	0.32	0.42	0.37	0.32	0.41
undefined:DP-SG	-0.75	-0.83	-0.67	-0.75	-0.83	-0.67
false:subject-final	0	-0.03	0.03	0	-0.03	0.03
undefined:subject-final	0	-0.03	0.04	0	-0.03	0.04
DP-SG:subject-final	0	-0.02	0.02	0	-0.02	0.02
false:DP-SG:subject-final	0.01	-0.01	0.04	0.01	-0.01	0.04
undefined:DP-SG:subject-final	-0.01	-0.04	0.03	-0.01	-0.04	0.03
SD-false	-0.18	-0.24	-0.13	-0.18	-0.24	-0.13
SD-undefined	-0.1	-0.15	-0.05	-0.1	-0.15	-0.05
SD-DP-SG	-0.22	-0.26	-0.18	-0.22	-0.26	-0.18
SD-subject-final	0.03	-0.01	0.06	0.02	-0.01	0.06
SD-false:DP-SG	0.18	0.13	0.23	0.18	0.13	0.23
SD-undefined:DP-SG	-0.29	-0.34	-0.23	-0.28	-0.34	-0.23
SD-false:subject-final	0.03	-0.02	0.09	0.03	-0.02	0.08
SD-undefined:subject-final	-0.08	-0.13	-0.02	-0.08	-0.13	-0.02
SD-DP-SG:subject-final	-0.01	-0.05	0.03	-0.01	-0.05	0.03
SD-false:DP-SG:subject-final	0.1	0.05	0.16	0.1	0.05	0.15
SD-undefined:DP-SG:subject-final	-0.16	-0.22	-0.11	-0.16	-0.22	-0.11

Table 3: Model summary for the critical items in Experiment 2. All $\hat{R} = 1$ and all effective sample sizes for bulk and tail $> 10k$. Standard deviation parameters are prefixed with 'SD-' and are reported on the log scale.

Parameter	ETI Median	$ETI_{2.5}$	$ETI_{97.5}$	HDI Mode	$HDI_{2.5}$	$HDI_{97.5}$
Intercept	1.54	1.34	1.74	1.54	1.34	1.75
SD-Intercept	-0.51	-0.6	-0.41	-0.51	-0.6	-0.41
doublepl	-0.97	-1.28	-0.65	-0.97	-1.28	-0.66
indefundef	-3.27	-3.49	-3.05	-3.29	-3.48	-3.04
relatumundef	-3.37	-3.56	-3.17	-3.38	-3.57	-3.18
relatumundefpl	-3.35	-3.55	-3.15	-3.36	-3.56	-3.16
scalarfalse	-0.96	-1.21	-0.71	-0.95	-1.21	-0.71
uniquenessviol	-0.83	-1.16	-0.5	-0.82	-1.15	-0.49
subject-initial	0.06	-0.08	0.19	0.06	-0.08	0.19
doublepl:subject-initial	-0.06	-0.29	0.18	-0.06	-0.29	0.18
indefundef:subject-initial	-0.12	-0.28	0.05	-0.12	-0.28	0.05
relatumundef:subject-initial	-0.03	-0.2	0.14	-0.03	-0.21	0.13
relatumundefpl:subject-initial	-0.04	-0.2	0.12	-0.04	-0.2	0.12
scalarfalse:subject-initial	-0.11	-0.33	0.11	-0.12	-0.34	0.11
uniquenessviol:subject-initial	-0	-0.24	0.23	-0.01	-0.25	0.22
SD-doublepl	0.17	0.03	0.31	0.18	0.02	0.31
SD-indefundef	-0.3	-0.45	-0.16	-0.3	-0.45	-0.16
SD-relatumundef	-0.68	-0.82	-0.52	-0.67	-0.83	-0.52
SD-relatumundefpl	-0.66	-0.8	-0.51	-0.66	-0.81	-0.51
SD-scalarfalse	0.06	-0.08	0.21	0.07	-0.08	0.22
SD-uniquenessviol	0.04	-0.11	0.19	0.04	-0.11	0.19
SD-subject-initial	-0.07	-0.2	0.05	-0.07	-0.2	0.05
SD-doublepl:subject-initial	-0.06	-0.26	0.14	-0.05	-0.26	0.14
SD-indefundef:subject-initial	-0.44	-0.64	-0.24	-0.43	-0.63	-0.23
SD-relatumundef:subject-initial	0.38	0.18	0.58	0.38	0.19	0.58
SD-relatumundefpl:subject-initial	-0.09	-0.29	0.11	-0.1	-0.29	0.11
SD-scalarfalse:subject-initial	-0.14	-0.34	0.07	-0.14	-0.34	0.07
SD-uniquenessviol:subject-initial	0.07	-0.13	0.27	0.07	-0.13	0.27

Table 4: Model summary for the control items in Experiment 2. All $\hat{R} = 1$ and all effective sample sizes for bulk and tail $> 10k$. Standard deviation parameters are prefixed with 'SD-' and are reported on the log scale.

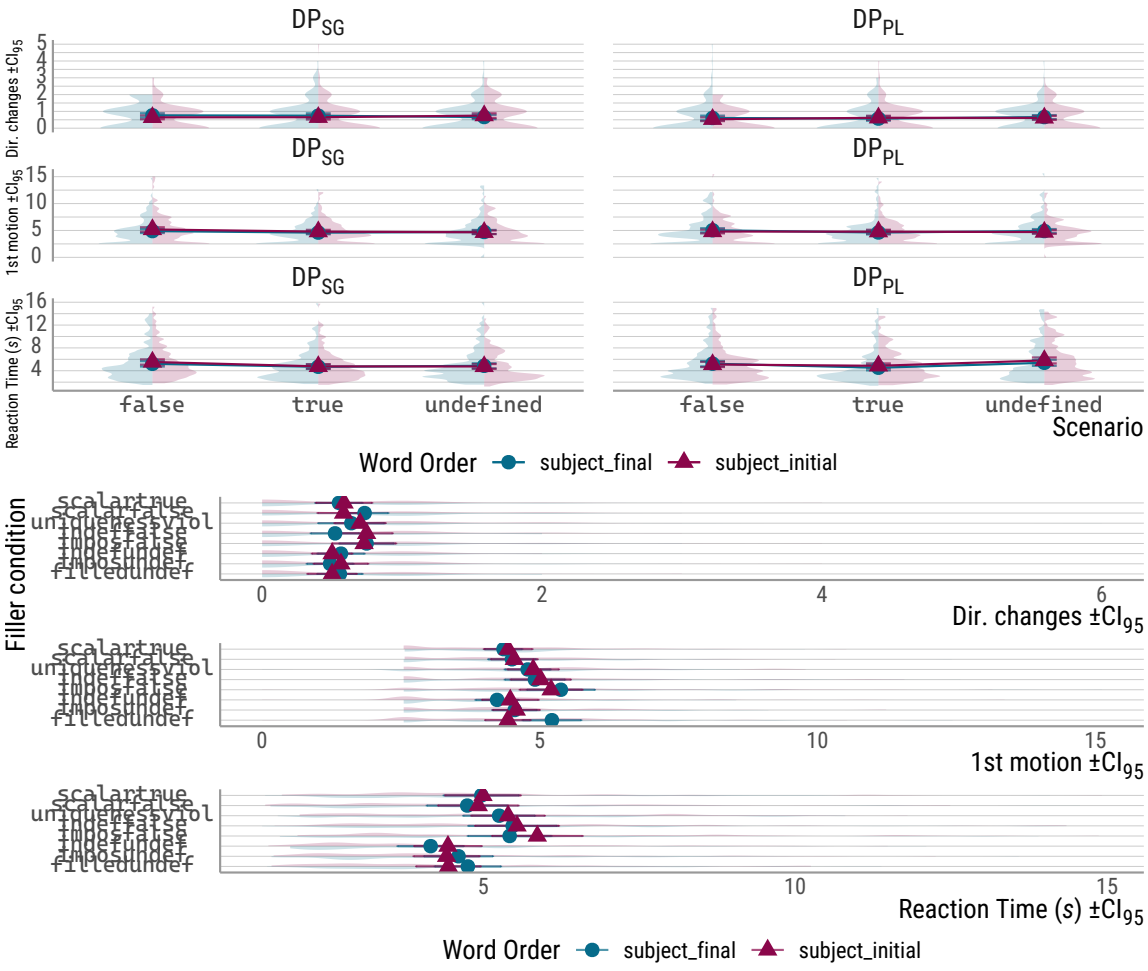


Figure 12: Online measures associated with the critical and the control items of experiment 1.

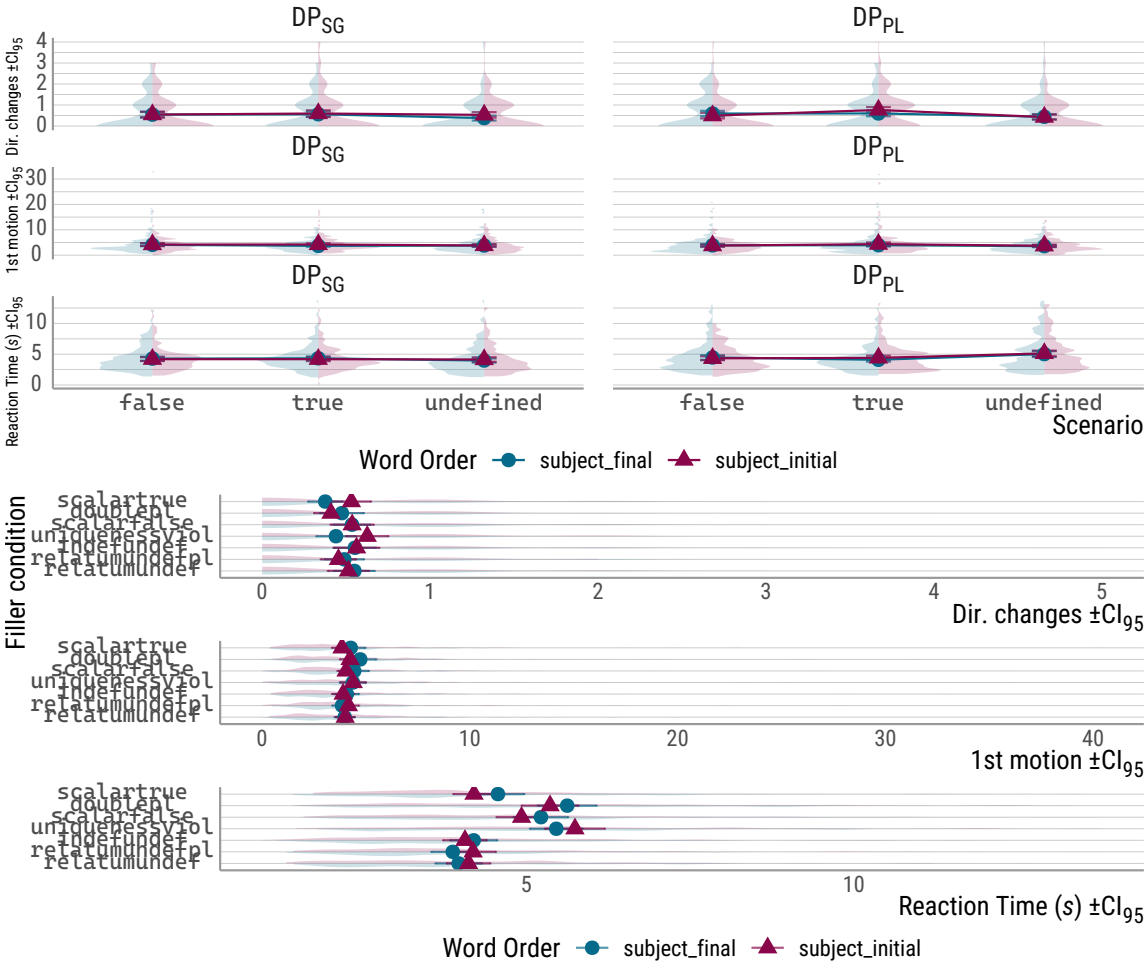


Figure 13: Online measures associated with the critical and the control items of experiment 2.

A Model tables

A.1 Experiment 1

A.2 Experiment 2

B More measures

B.1 Experiment 1

B.2 Experiment 2

C By-participant plots

C.1 Experiment 1

C.2 Experiment 2

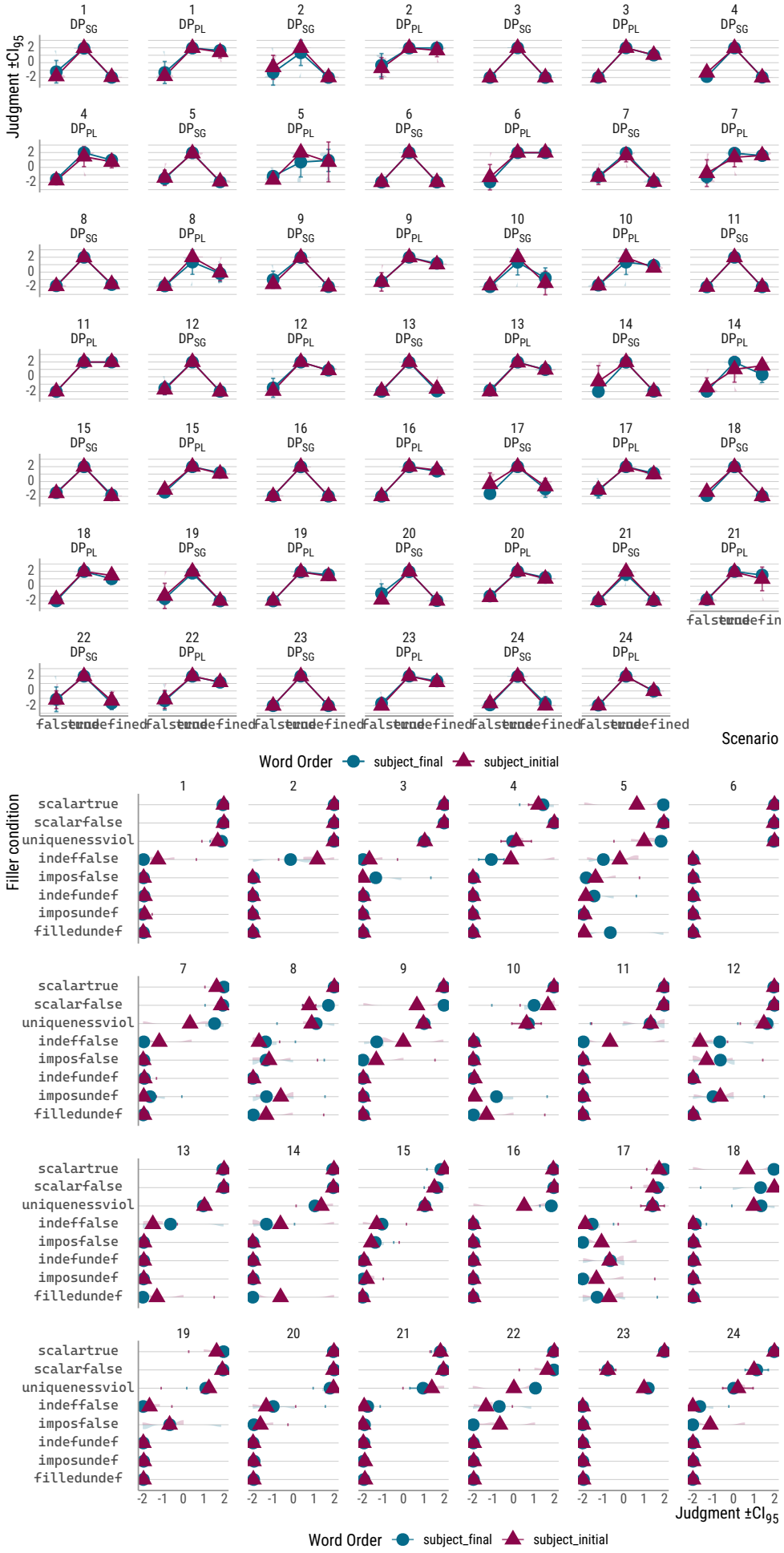


Figure 14: By-participant plots for the critical (*top*) and control (*bottom*) portions of experiment 1.

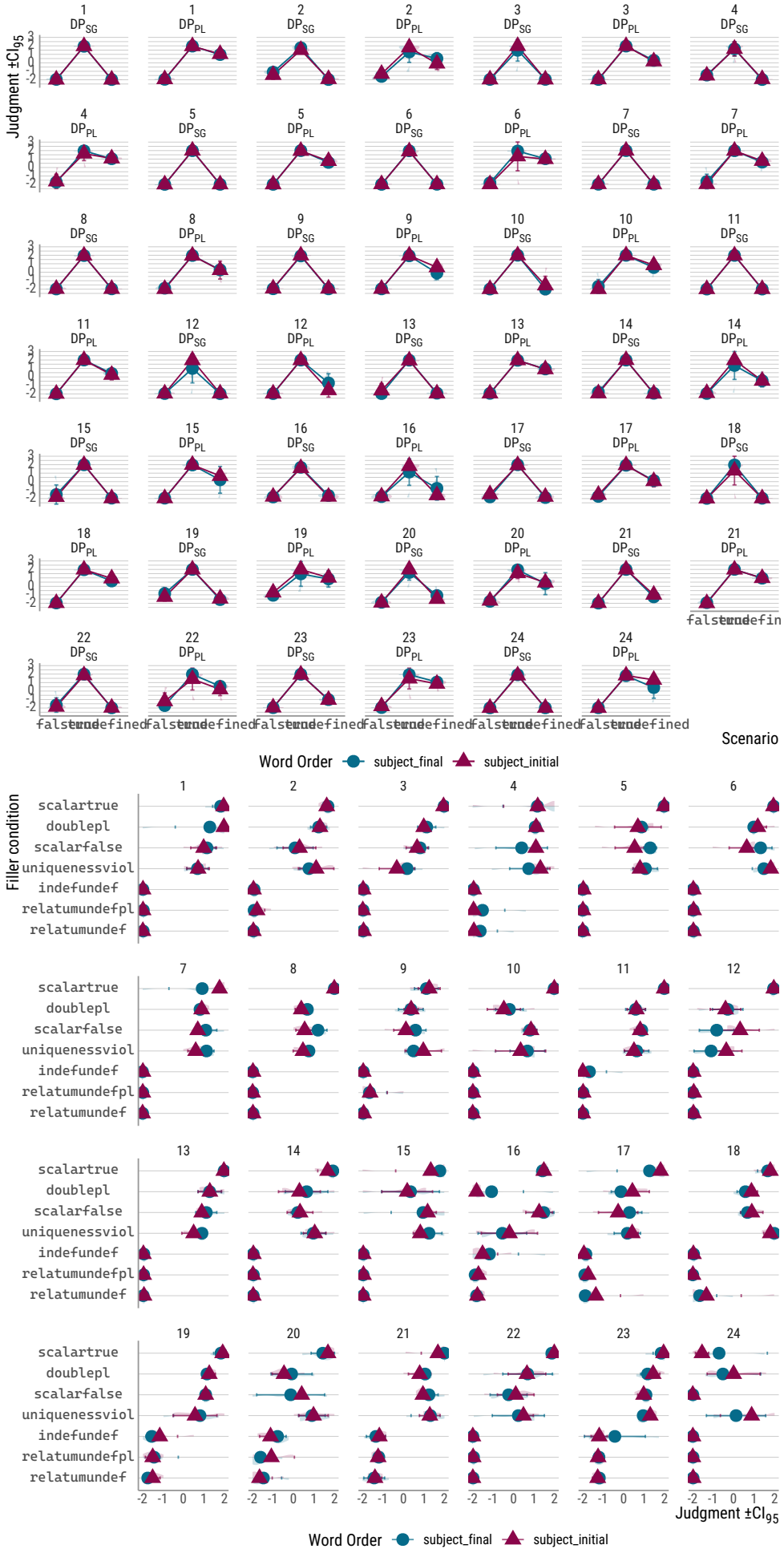


Figure 15: By-participant plots for the critical (*top*) and control (*bottom*) portions of experiment 2.