Slowdowns in scalar implicature processing: Isolating the intention-reading costs in the Bott & Noveck task

Camilo R Ronderos, Ira Noveck

Slowdowns in scalar implicature processing: Isolating the intention-reading costs in the Bott & Noveck task

Camilo R. Ronderos[1] & Ira Noveck[2]

[1] University of Oslo, Department of Philosophy, Classics, History of Art and Ideas

[2] Laboratoire de Linguistique Formelle, UMR 7110 CNRS-Université de Paris

Author notes.

The authors contributed equally. The authors made the following contributions. Camilo R. Ronderos: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing, Formal Analysis, Visualization, Data curation, Methodology, Software; Ira Noveck: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing, Supervision, Funding acquisition, Project administration, Formal Analysis, Methodology.

This work was supported by a post-doc grant through Oslo University to the first author and a Chaire d'Excellence grant from the *Université de Paris-Cité* to the second author. Correspondence concerning this article should be addressed to Ira Noveck, LLF, CNRS – UMR 7110 Université Paris Diderot-Paris 7. E-mail: ira-andrew.noveck@cnrs.fr

The authors thank Kira van Voorhees for her contributions in setting up Experiment 1.

Abstract

An underinformative sentence, such as *Some cats are mammals*, is trivially true with a semantic (*some and perhaps all*) reading of the quantifier and false with a pragmatic (*some but not all*) one, with the latter reliably resulting in longer response times than the former in a truth evaluation task (Bott & Noveck, 2004). Most analyses attribute these prolonged reaction times, or costs, to the steps associated with the derivation of the scalar implicature. In the present work we investigate, across three experiments, whether such slowdowns can be attributed (at least partly) to the participant's need to adjust to the speaker's informative intention. In Experiment 1, we designed a web-based version of Bott & Noveck's (2004) laboratory task that would most reliably provide its classic results. In Experiment 2 we found that over the course of an experimental session, participants' pragmatic responses to underinformative sentences are initially reliably long and ultimately comparable to response times of logical interpretations to the same sentences. Such results cannot readily be explained by assuming that implicature derivation is a consistent source of processing effort. In Experiment 3, we further tested our account by examining how response times change as a function of the number of people said to produce the critical utterances. When participants are introduced (via a photo and description) to a single 'speaker', the results are similar to those found in Experiment 2. However, when they are introduced to two 'speakers', with the second 'speaker' appearing midway (after five encounters with underinformative items), we found a significant uptick in pragmatic response latencies to the underinformative item right after participants' meet their second speaker (i.e. at their sixth encounter with an underinformative item). Overall, we interpret these results as suggesting that at least part of the cost typically attributed to the derivation of a scalar implicature is actually a consequence of how participants think about the informative intentions of the person producing the underinformative sentences.

Keywords: scalar implicatures, pragmatic processing, experimental pragmatics, mindreading
Word count: 10500

Slowdowns in scalar implicature processing: Isolating the intention-reading costs in the Bott & Noveck task

Introduction

Scalar implicature is arguably the most studied phenomenon in the experimental pragmatic literature. Analyses typically center on a particular word in a sentence, such as *some*, and how it is implicitly enriched (e.g., to mean *not all*) when expressed as part of an utterance. This *pragmatic* enrichment justifies false evaluations of sentences such as (1):

(1) Some cats are mammals.

This is not the only way to interpret (1). The *semantic* meaning of *some* is compatible with *all*; thus, one is justified to consider (1) as trivially true when it is assumed to arrive without the pragmatic enrichment. One of the ongoing debates about scalar implicatures concerns the processing costs associated with the pragmatic enrichment. As research in this area has grown, two opposing views have emerged on the matter.

On one side is a preponderance of evidence showing that pragmatic enrichments are associated with effortful processing. Bott and Noveck (2004; from here on, *B&N*), the first study to make this claim, used sentences just like the one in (1). The authors required participants to evaluate such propositions (as TRUE or FALSE) along with control items such as those in (2-6):

(2) Some mammals are cats.
(3) Some cats are insects.
(4) All cats are mammals.
(5) All mammals are cats.
(6) All cats are insects.

Two general outcomes are associated with this experiment. First, participants are roughly equivocal in responding TRUE versus FALSE to cases like (1), which we will refer to as T1 (for Type 1) sentences as per B&N (likewise, the control items will be referred to as T2-T6). Second, response times to T1 items that yield TRUE responses (those relying on semantic, or *logical*, interpretations of *some*) are generally comparable to those of control items, whereas response

times to T1 items that yield FALSE responses (those relying on pragmatic interpretations) take much longer overall, at least a half a second longer on average than the logical responses. These findings made *B&N* a landmark reference for arguing that a scalar enrichment is an effort-demanding process that brings about a comprehension cost for language users (see also Noveck & Posada, 2003). B&N's effortfulness claims eventually generalized to other tasks using different paradigms (e.g., Breheny et al., 2006; Bott et al., 2012; De Neys & Schaeken, 2007; Dieussaert et al., 2011; Heyman & Schaeken, 2015; Khorsheed et al., 2022; Marty & Chemla, 2013; Rees & Bott, 2018; van Tiel et al., 2018; van Tiel et al., 2019; see Noveck, 2018; Breheny, 2019; or Khorsheed et al., 2022 for reviews). Most often, the source of the effortful pragmatic process is attributed to a combination of two steps: (a) the production of an alternative proposition (e.g. an utterance with *some* is said to generate a proposition with an alternative quantifier such as *all*) which is justified by some feature of context and (b) the negation of the alternative (e.g., see Bott and Frisson, 2022).

On the other side of the debate are a handful of studies showing that processing costs related to pragmatic readings can be shown to be minimal or even non-existent. That is, apparent costs that were heralded in early scalar implicature tasks (a) can be reduced or made to disappear (Degen & Tanenhaus, 2015; Foppolo & Marelli, 2018; Grodner et al., 2010), (b) can be manipulated as a function of a participant's knowledge state vis a vis the speaker (Breheny et al., 2013) or (c) shown to not arise immediately when a reader encounters a trigger word as one would expect in a vignette (Politzer-Ahles & Fiorentino, 2013). Typically, these studies point to features of context (e.g. something that makes an alternative more salient) or of presentation (how critical words are expressed) that mitigate processing costs. For example, Grodner et al (2010), managed to reduce previously reported processing costs and to get more immediate reactions to critical terms by, among other things, using the expression *summa* instead of *some of*.

Given the seemingly conflicting literature, can one -- or can one not -- claim that deriving a scalar implicature comes at a cost? The current work takes on this bivalent question and, over the course of the paper, considers a third possibility, which is that processing costs in scalar implicature tasks are present but transient. We argue, based on Paul Grice's (1989) work, that at least part of the processing cost associated with deriving a pragmatic response is linked to the participant's attempt to infer a speaker's intention, and that this cost is eventually reducible. To elucidate this point, we first take a closer look at Grice's influential account of communication

before turning to the B&N task, in which we aim to uncover intention-reading costs related to its well-known pragmatic readings.

Revisiting Grice

Part of the appeal of the two-step model of implicature derivation is that it resonates with the proposal by Grice (1989) concerning his Conversational Principle, according to which members of a cooperative exchange keep their contributions truthful and informative, relevant and lucid (as readers of this literature know, these features refer respectively to Grice's maxims of Quality, Quantity, Relevance and Manner). Importantly, Grice proposed that implicit meanings arise through conversational implicatures when an utterance violates a maxim.

Grice's pioneering account comports with current explanations of scalar implicature slowdowns because both involve detecting, exploiting, or enriching an underinformative (or ambiguous) sentence in order to generate a more informative alternative. However, most current mechanistic accounts of scalar implicature are not viewed through the larger lens of Grice's theory, which assigns a prominent role to working out the speaker's intentions during conversation. According to Grice, communication is successful when a speaker, who has the intention of sharing some belief with an addressee, gets their addressee to recognize the speaker's intention as well as to recognize that the speaker intends to share it. In other words, a successful speaker shares an *informative* intention so that the addressee acquires a new belief; this is then embedded in a higher-order *communicative* intention that makes manifest the speaker's informative intention (Scott-Phillips, 2008; Sperber & Wilson, 1986; see also Helming et al., 2016; Wilson & Sperber, 2004). Note that this is not how Gricean theory is usually viewed in the experimental linguistic-pragmatic literature. Most applications of Grice's theory (and their descendants) analyze a sentence and draw out implicit information by mechanically applying the maxims and inferential rules. Such steps may well occur, but the pre-eminent role of working out the speaker's intentions in the process is usually left out.

To formulate it in more current psychological terms, it can be said that the Gricean account calls for addressees to consider a speaker's mental states (specifically, their intentions and beliefs) when interpreting their utterances (e.g., Grice, 1989, p. 123). Such a process falls under several banners, such as *mindreading* (Nichols & Stich, 2003; Spaulding, 2020) or *mentalizing* (Frith & Frith, 2006) and is part of literatures referred to as *Theory of Mind* (Baron-

Cohen, Leslie, & Frith, 1985) or *Folk Psychology* (Breheny, 2006). With respect to inferential meaning, it is plausible that the processing cost associated with deriving a scalar implicature, especially of the kind reported in *B&N*, is not uniquely associated with the mechanics behind the enrichment of *Some* to *not all* but with the way in which participants in an experiment attempt to understand the intention of the given utterance's producer (which we refer throughout the manuscript as *mindreading*). This kind of Gricean analysis has been successfully applied to understanding the processing of other pragmatic phenomena (e.g., the comprehension of irony); in our experiments, we investigate the extent to which such an analysis can be applied to cases of scalar implicature.

In the remainder of this Introduction we take the following three steps. First, we broadly review the experimental literature on scalar implicatures in order to make a distinction between, on the one hand, studies that measure latencies while participants evaluate equivocal statements (e.g., *B&N*'s T1 sentences) and, on the other, studies that are concerned with the speed of processing of sentences (or parts of them) that are made by presumably reliable speakers. Second, we consider other experimental pragmatic phenomena in which a speaker's intention is crucial to latency measures, viz. irony-processing (Spotorno & Noveck, 2014), precedent-uses in reference-making (Kronmüller & Barr, 2015) and contrastive inference among reliable and unreliable speakers (Grodner & Sedivy, 2011). Results from these areas provide relevant clues for the work we pursue here. Finally, we describe the reasoning behind the evolution of our three experiments; the latter two of which provide evidence showing that intention reading itself plays an overlooked and critical role in pragmatic response slowdowns on the *B&N* task.

Different tasks, different outcomes?

B&N's results led many to investigate the time course of responses that are associated with pragmatic versus logical readings of sentences that use weak, underinformative scalar terms. These follow-up studies were important because, in most cases, they were determined to test whether the pragmatic/logical response time difference found in *B&N* held as tasks became more ecologically valid. We describe two of the early (and seminal) offshoots here. In one, Breheny, Katsos, and Williams (2006) investigated participants' reading times of phrases in the context of short vignettes. For example, in their first experiment, they presented disjunctive phrases that are logically compatible with conjunctions but pragmatically distinguishable from them when understood as exclusive. Imagine reading about a character, John, who while studying for exams,

relied on *the class notes or the summary*. When inserted in a context that invites an upper-bound (exclusive) reading (e.g. the participant reads that John did not have much time so he needed to make a decision about what means he would use to study), the disjunctive phrase took reliably longer to read than when placed in a context inviting a lower-bound (inclusive) reading (in which the participant reads that John was desperate to pass a difficult course).

Now, consider the line that is based on eye-tracking from Huang and Snedeker (2009). They showed that the word *some*, in a sentence such as *Point to the girl that has some of the socks*, did not automatically prompt looks to a target (i.e. one of two girls depicted in a scenario that also included two boys) when one girl had two of the scenario's four socks and the other girl had all four of the scenario's soccer balls (note how the critical words *socks* and *soccer balls* prompt ambiguity at the outset). Participants generally needed to hear the entirety of the pluralized noun to know which girl to target and click on, indicating that both targets (one referentially linked to *Some* and another to *All*) were initially compatible with each other. Collectively, both classes of studies, along with many others using other techniques (e.g. Bott et al., 2012; De Neys & Schaeken, 2007; Khorsheed et al., 2022; Marty & Chemla, 2013; Tomlinson et al., 2013; Tomlinson & Ronderos, 2021; Van Tiel & Schaeken, 2017), lent credence to the notion that readings that include a weak scalar term do not automatically trigger enriched meanings and that narrowed, enriched readings are likely associated with supplementary processing when compared to those without such interpretations.

These sorts of confirmatory findings were eventually questioned by studies that did not routinely find slowdowns linked to underinformative scalar utterances. The studies that challenged the effortfulness claims were typically those that employed a Visual World (eye-tracking) Paradigm. For example, Grodner et al (2010) – besides changing the presentation of *some of* to *summa* – made several modifications to render salient a *some but not all* reading in the Huang & Snedeker task: They also removed numerical controls (Huang & Snedeker, 2009 had included items such as *Point to the girl that has two of the socks*), they added more targets and made the distribution of articles (e.g. of soccer balls and socks) more explicit in two steps before the arrival of the test sentences. With these modifications, looks to targets in reaction to *Point to the girl that had some of the soccer balls* were inexorably linked to cases in which she had two of four soccer balls and at a speed that was comparable to looks in reaction to *Point to the girl that had all of the socks*. Data like these undermined claims that said that the working out a scalar

implicature is universally time-consuming. With particular constraints, one could coax participants to apply a *some-but-not-all* reading more readily.

When the task involves truth-evaluations à la B&N, processing cost reductions are less apparent. That is, it is generally accepted that B&N tasks reliably produce slowdowns linked to pragmatic responses to statements like the one in (1) (van Tiel & Schaeken, 2017). One near exception is from Degen & Tanenhaus (2015, Experiment 3) which was a reaction time study that included a form of truth-evaluation (*Yes they agree with* or *No they disagree with* a spoken description) in a gumball counting paradigm (in which gumballs fall from one compartment to another and where *all*, *some* or *none* of them fall from one to the next). The authors failed to find significant effects linking effortful pragmatic processing to *No* responses in one model (among participants who gave completely consistent responses within either the *some* or the *summa* condition for conditions whose outcomes correspond to *all*). As the authors noted, however, their paradigm produced an overwhelming majority of semantic responders (81%); i.e. only 7 of 47 participants could be said to be pragmatic, making this instantiation of a truth-evaluation task unlike others that evaluate the speed of pragmatic judgements.

As this brief review of the literature reveals, questions remain concerning the costs associated with scalar implicature derivation. Findings from Visual World experiments challenge the claim that pragmatic readings of scalar implicature *consistently* come at a cost and, in light of such data, it is necessary to develop an account that provides a principled explanation for the variability of processing costs reported across scalar implicature tasks. This is what we aim to do in the current manuscript by investigating the role of intention-reading behind the processing cost typically found when using the B&N task.

Intention-reading in Bott & Noveck (2004)

That there is likely more to the B&N effects than just the enrichment of *some* to *not all* can be appreciated when considering the task in greater detail. We begin by pointing out that the *B&N* task presents a (faceless, voiceless and) inconsistent "speaker" who makes statements that are patently true sometimes (e.g., *Some mammals are elephants, All elephants are mammals*) and patently false the rest of the time (e.g., *Some elephants are insects*). It also includes one type of item that can have two readings (the T1 items). As such, the speaker's informative intention is likely to be perceived in at least one of two ways by an audience. One is to understand T1 as trivially true. The other is to understand it as false, which is possible if one makes adjustments to

what was literally said. To put this in Gricean terms (which, again, can be encapsulated by the idea that communication is successful when the listener recognizes a speaker's communicative and informative intentions), we argue that the addressee (the participant) is made aware of the speaker's communicative intention in the task from the start, namely when the experiment requests participants to make true/false judgements for each item. When presented with a T1 item (which has two possible readings), there will be those who will assume that the speaker's informative intention is to convey that this item is false. It is for these participants that the pragmatic adjustment is justified.

This leads us to spell out the following argument in three steps. First, given that sentences have meaning in as much as they are used, participants naturally assume that the sentences are produced by someone (a 'speaker'). Second, for those participants who respond FALSE to a T1 sentence in the B&N paradigm, they are not simply reading the sentence bottom up and determining that it is false because, say, the quantifier was underinformative and prompted a maxim violation, or because the word *some* eventually generates a *not all* reading. Rather, they are determining what it is that the Experiment's anonymous speaker intends to communicate. Third, in contrast to those who draw a logical reading of a T1 item, those who respond FALSE reason that it is the speaker's intention to convey that this item is false. The last step arguably plays an important part in the slowdowns.

According to our view, the making of a *false* attribution of a T1 item takes place with respect to a speaker and not to the statement. Based on past experiments with the B&N task, one can assume that roughly half of participants understand that the informative intention behind the presentation of a T1 item is akin to *the speaker of this item wants me to believe that it is false*. Importantly, the processing profile for arriving at this informative intention is different than one that assumes a re-occurring mechanical two-step enrichment with each encounter with *some*. That is, once a pragmatic participant establishes what they consider to be a speaker's informative intention with respect to T1 items, they need not repeat the intention-reading steps each time they encounter a T1 sentence and interpret it pragmatically. This last claim motivates our Experiments 2 and 3.

Pragmatic processing effects and mindreading

Part of our approach is to consider a wider array of experimental pragmatic phenomena and especially those that more obviously rely on a participant's intention-reading. We specifically

turn to three experimental pragmatic areas that less controversially rely on intention-reading -- irony-processing, referential precedents when naming novel objects, and contrastive inference delivered by reliable and unreliable speakers. Each is revealing of speaker-related effects that resonate with ones we assume arise in the B&N task.

A classic finding in the irony processing literature is that out-of-the-blue ironic utterances take longer to process than their literal equivalents (e.g., Filik & Moxey, 2010; Giora & Fein, 1999). It has nevertheless been shown that this difference can be mitigated through various kinds of manipulations. One intriguing way to reduce differences over the course of an experiment is to -- practically predictably -- present vignettes whose ironic sentence consistently arise in the wake of a vignette's negative event. Under such conditions, slowdowns of ironic statements (relative to literal controls) are only noticeable in the early part of an experimental session. Spotorno and Noveck (2014) argued that such *early-late effects* are a consequence of participants engaging in their mindreading abilities: Once participants understand how to interpret potentially incongruous remarks early in a session (ostensibly from a single narrator who presents repeatable types of items), their speed in understanding the intent behind ironic sentences speeds up to the point that they are read as fast as literal controls by the end of the experimental session (see also Olkoniemi, Ranta, and Kaakinen 2016, for a related finding). Following up on this study, Ronderos, Tomlinson and Noveck (2023) found that explicitly manipulating the degree to which a participant could anticipate a speaker's intention affected the processing effort of target ironic sentences, to the point that they could be *faster* at reading ironic responses compared to their literal controls. The more recent studies further support the idea that the degree to which mindreading engagement is facilitated (via the repeated appearance of ironic sentences or by generating strong expectations of a speaker's informative intention) influences (i.e. reduces) the processing costs that are required to understand an ironic remark.

Another area of Experimental Pragmatics that considers the processing costs related to recognizing a speaker's intention concerns the establishment of referential conventions, a quintessential pragmatic undertaking that began with the work of Clark & Wilkes-Gibbs (1986). To make this experimental paradigm concrete, consider a speaker who labels one unusual object out of three so that a participant/listener can find it. By so doing, the speaker introduces a *referential precedent* into the discourse and reduces ambiguity for the purposes of an ongoing conversation (Kronmuller & Noveck, 2019). Now imagine that a second labeler appears on the scene. That person either comes up with their own name for the unusual object or, by chance,

uses the same one that the previous labeler used. For the sake of completeness, imagine too that the original labeler suddenly attributes a new name for a previously named object. Once one considers speaker's identity as crucial to reference (Brown-Schmidt, 2009), three effects emerge (see also Pogue et al, 2016, for further evidence of speaker-specific effects on deriving pragmatic inferences). Two of these occur, not surprisingly, when the original speaker, who created a precedent for an object, suddenly comes up with a new name for it. The eye-tracking data show that name switching from a single speaker leads to momentary confusion on the part of the participant both when compared to i) cases in which a previously coined reference is *repeated* by the same speaker as well as to; ii) cases in which a new speaker comes up with a new label. That is, hearing a new name for a previously named object is not viewed as unusual when it comes from a new labeler. The third effect is more surprising: the data also show that if a newly appearing speaker were to use the same label as a previous interlocutor, it also leads to a slight slowdown when compared to the case in which there is a single interlocutor. This paradigm sensitively shows how participants routinely consider a speaker's informative intention. A meta-analysis from Kronmüller and Barr (2015) incorporating 10 studies confirms these effects, even though the last is far weaker than the two others.

Two studies on contrastive inference effects bolster the claim that participants rely on speaker-specific information when establishing reference. Grodner and Sedivy (2011) and Gardner et al. (2021) found that when participants believe a speaker to be reliable, they rapidly derive contrastive inferences when hearing sentences with scalar adjectives such as *click on the large cup* (in other words, there are early looks to the target item when there is a juxtaposed contrasting object, such as small cup) confirming previous findings (Sedivy, 2003; Sedivy et al., 1999). However, when the speaker is believed to be unreliable, participants are less likely to generate such inferences. As a possible explanation for the finding, Gardner et al. (2021) suggest that comprehenders reason about a perceived unreliability (e.g., *Why did a speaker produce "large" in a given context?*) and use this as a data point to make predictions for interpreting subsequent input. If such an unreliability is attributed to a speaker's idiosyncrasy, the predictions point to speaker effects. This is consistent with our claim that determining an individual speaker's intentions is an important part of processing instances of pragmatic language use.

Taken together, the findings reported in this section indicate that participants are sensitive to the fact that message-deliverers are sharing intentions and that part of a participant's task is to identify what these are. We take advantage of speaker-specific observations concerning the first

two of these experimental pragmatic literatures – on irony processing and referential precedents -- as we aim to show that intention-reading with regard to the speaker applies equally to the *B&N* task.

The strategy behind our investigation

In this work, we revisit *B&N*'s paradigm in order to better determine what is behind the processing cost of pragmatic readings of its underinformative (T1) sentences, with the hypothesis that an important part of this cost is related to informative intention-reading. Given that there were in fact several variations of the *B&N* task across its four experiments, our first step consisted in carrying out an experiment whose goal was to arrive at a web-based version of the *B&N* task that most reliably provides the kind of distributions (roughly equal amounts of pragmatic and logical readings of T1s) and the typical slowdowns linked to pragmatic responses, before introducing our manipulations in Experiments 2 and 3. That is, *B&N* included (a) tasks whose target sentences were presented one word at a time or else as full sentences as well as; (b) tasks that rendered critical the available time to read the test sentences (e.g. see *B&N*'s Experiment 4 which showed that extra time was associated with more pragmatic responding). We therefore decided to manipulate these two parameters in order to find what we call the *pragmatic sweet spot*: the optimal online experimental setting for uncovering *B&N*'s well-known effects. We did this while introducing two other more general changes (compared to B&N's original study): We tested native English speakers (the original was conducted in French though the B&N paradigm has since been tested in English) and, as we intimated, we adapted the paradigm to a web-based format.

Our second step was to test our claim that the processing difference between logical and pragmatic readings of under-informative T1 sentences is due, at least in part, to a participant's effort to read the intentions of a task's implicit speaker. We thus anticipate finding results that are consistent with those found in neighboring literatures that more obviously rely on pragmatic readings. Much like with Spotorno & Noveck's (2014) *early-late* effects, we similarly predict these for the B&N task. The more a comprehender anticipates a pragmatic speaker's informative intention, the less costly it should be to derive the pragmatic response as the task wears on. It follows that one should find that the classic extra time associated with pragmatic (FALSE) responses to T1s (which we take to be a holistic measure of processing effort) should diminish over the course of the experimental session with repeated pragmatic readings of the same class of

sentences. By the end of the session, the speed of pragmatic interpretations ought to elicit processing costs that are comparable to those required for logical interpretations. This means that an isolable amount of the extra-processing linked to pragmatic responses to T1 sentences should be evident in the early trials (thus producing *early-late* effects). To be more precise, we predict that the latencies of pragmatic responses should be relatively slow with respect to semantic responses early in an experimental session. This would be a novel result for this literature, and it is the pre-registered prediction at the core of Experiment 2.

Finally, Experiment 3 tests the degree to which these potential early-late effects of scalar implicature processing are speaker-specific. If an important part of the processing effort traditionally associated with the derivation of an implicature in the *B&N* task is a consequence of determining the intention of an individual speaker, then switching the speaker midway through the experiment should affect these early-late effects. Introducing a new speaker should re-establish the need for participants to determine the speaker's informative intentions. Experiment 3's manipulation resonates with those found in the literature on referential precedents. Experiments 2 and 3 were pre-registered. All data, analysis scripts, pre-registrations and materials can be found on the project's OSF repository:
https://osf.io/msjcq/?view_only=5d4937f944514bf98e72f36f9756c74f

Experiment 1: Finding the 'pragmatic sweet spot'

Participants

216 participants were recruited via Prolific. These were monolingual, right-handed, native speakers of British English currently residing in the UK. They were all between 18-35 years of age. They each received £1.25 as compensation for participating in the study.

Materials

We translated the *B&N* materials to English. These consisted of six types of statements, as shown in examples T1-T6 above. The statements belonged to six categories (insects, mammals, reptiles, fish, shellfish, and birds) with nine category-members for each, which amounts to a total of 54 items. We modified some of the items that we surmised were less evident to English speakers. For example, *langoustines* was replaced with *barnacles*. Moreover, since every category-member in the original French list had a plural marker, we presented category-member items in plural if possible (e.g., we used *shrimps* instead of *shrimp*).

Design

Our experiment included the original 6 conditions from *B&N*, illustrated in Examples T1-T6. Additionally, we manipulated two orthogonal dimensions between participants. One was the delay (from 0 to 1 to 2 seconds) between sentence onset and the availability of the *True* and *False* options on the screen. The other manipulation concerned the presentation of the stimuli, which could be as a full sentence or else as one-word-at-a-time. The combinations of the factors DELAY and SENTENCE PRESENTATION amount to six different experimental variations. Each of these variations incorporated a latin-squared design for the factor STRATEGY: The six sentence types of each item were distributed across 6 lists. This amounts to a total of 36 experimental lists.

Procedure

The experiment was programmed using the Ibex experimental software (Drummond, 2013) paired with the PCIbex Farm (Zehr & Schwarz, 2018) and ran entirely online. Participants were told they would read sentences and would have to indicate whether they believed these sentences were TRUE or FALSE using their keyboard. Participants were shown a figure indicating how to keep their hands poised over their keyboard throughout the task and they were instructed to respond as quickly and accurately as possible. They were asked to complete the experiment in one sitting and to avoid distractions and interruptions. The experimental session was divided into three equally-spaced blocks.

Analysis

We analyzed the data in three ways. First, we counted the number of pragmatic and logical readings of critical sentences with respect to each experimental variation. Second, we counted the number of logical and pragmatic responders, operationalized as the number of participants who answered consistently in one direction for at least eight of the nine critical T1 trials (what we call *consistent responders*). Finally, we fitted mixed-effect linear models (including random intercepts and slopes by items and by participants for the factor STRATEGY) to the log-reaction times of each experimental variation. Our goal, based on B&N and the following literature using their paradigm, was to find the *pragmatic sweet spot*: The experimental variation that would (1) provide the most balanced number of pragmatic and logical readings of T1 sentences; (2) provide

the most balanced number of consistent pragmatic and logical responders and meanwhile; (3) find significantly longer response times for pragmatic relative to logical readings of T1.

Results

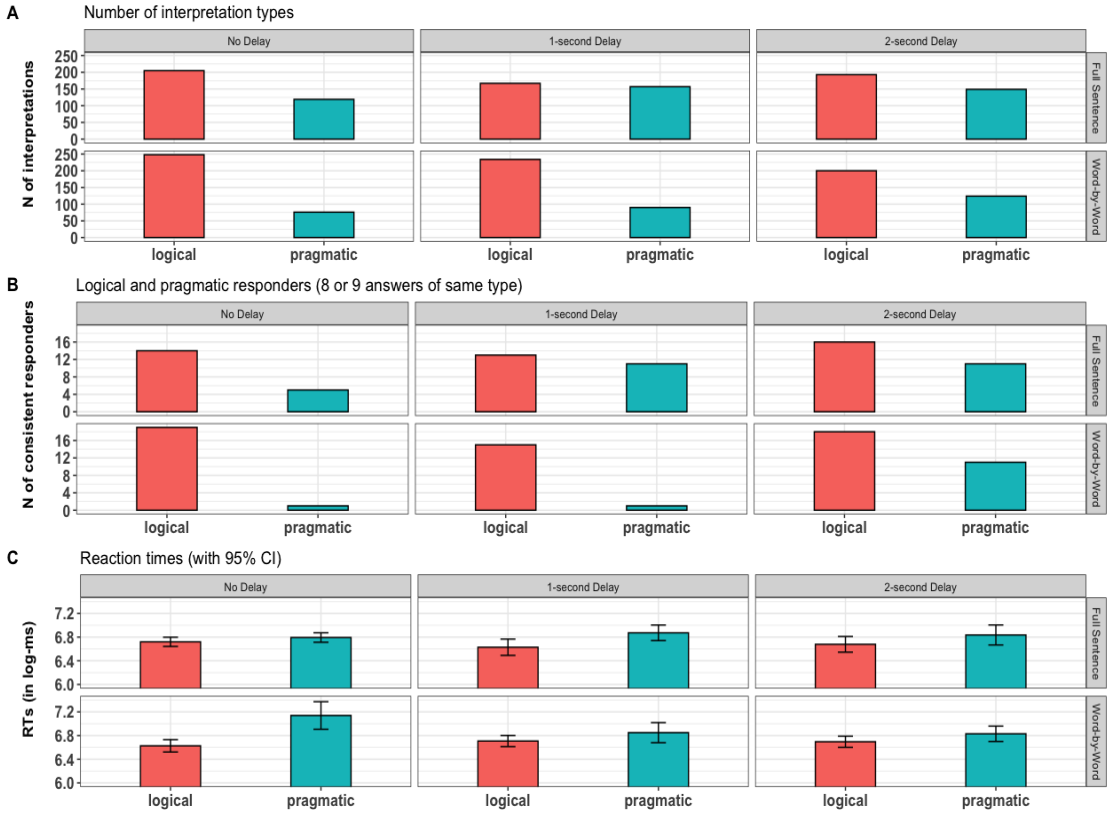

Figure 1. A summary of outcomes with respect to the T1 sentences: Total number of each type of interpretation, rates of consistent 'logical' and pragmatic responses and log-Reaction Times (with 95% confidence intervals).

The results are shown in Figure 1. Following our criteria, the experimental variation with a one-second delay and a 'full sentence' presentation type was deemed to represent the *pragmatic sweet spot*. In this variation, there were a total of 157 pragmatic and 167 logical responses (Panel B). Further, 11 participants replied consistently pragmatically and 13 consistently logically (Panel C). Finally, this variation found significantly longer log-RT's for pragmatic vs- logical responses (t=2.17, p<0.05) (Panel A). We therefore chose to use this variation as the basis for Experiments 2 and 3.

To test whether the 1-second delay caused a significant increase in pragmatic responses, we conducted an additional test. We first counted the number of pragmatic readings of T1s per participant in the 'full sentence' condition. We then fitted a simple linear model to this data with

DELAY as a predictor of number of pragmatic readings. We found that the 1-second delay significantly increased the number of pragmatic readings per participant relative to the 'no delay' condition ($t= 2.1$, $p<0.03$), whereas the 2-second delay did not ($t=1$, $p=0.27$).

Discussion

This study systematically varied two features that were inspired by manipulations in B&N's scalar implicature studies: the amount of time that participants are given before answering (0, 1 and 2 seconds) and the presentation of the experimental stimuli (as an entire sentence versus one-word-at-a-time). Our main goal was to identify an online version that would best reflect the findings reported in B&N's study. We found that this occurs when the sentence is presented in its entirety (like in B&N's Experiment 3) but with a one second delay from the moment that the test sentence is displayed. This is the condition that produced the greatest parity of logical and pragmatic responses while also producing a lag associated with the pragmatic response.

While we accomplished our goal of identifying a replicable version of the lab-run B&N task, we would be remiss if we did not underline three results from the *sweet spot* experiment that are edifying for the literature. One is that one can see (again) how an imposed delay (from immediate availability to respond versus 1 or 2 seconds) increases rates of pragmatic responding. This is consistent with the finding from B&N's Experiment 4, where participants were provided short or long response latencies and the latter were associated with higher rates of pragmatic responses. The second is that the *sweet spot* here does not line up with the original version of B&N's Experiment 3, which had the response options available immediately. As studies move increasingly to online environments, it is important to be aware that online-run experiments (obviously) come with less experimenter control which ultimately means that minor features can have an outsized impact on tasks and especially for those whose effects are time-sensitive. As far as our current goals are concerned, our caution paid off. That said, while B&N's Experiment 3 reported pragmatic responses at rates of around 60%, the equivalent condition in the current online study shows rates of pragmatic responding that were roughly half of that. The Experiment's third intriguing finding concerns the way that the one-word-at-a-time condition does not increase rates of pragmatic responding. When B&N ran their Experiment 4 (which forced one group of participants to answer within 900 msec and another within 3000 msec) they purposely chose the one-word-at-a-time method (instead of presenting the whole sentence at once) in order to make sure that participants read each of the words. Implicitly, B&N were

assuming that this innovation would allow for more thoughtful processing on the part of the participants and, if anything, for more pragmatic responding. Evidently, word-by-word processing here does not increase the likelihood of pragmatic responding.

Experiment 2: Trial effects and pragmatic processing

In the Introduction, we argued that pragmatic slowdowns are linked not only to linguistic structures that separate the pragmatic from the logical response, but also to intention-reading factors. Participants who respond false take into consideration that their interlocutor is equally likely to say true or false statements and they are able to get past the surface-level true reading of the T1 sentence in order to appreciate its potential for a false response. We argued that if we are right – that the cause of the pragmatic slowdown is due to mindreading concerning the anonymous speaker -- then one should find that the pragmatic reading to T1 sentences speeds up over the course of the study. That is, we predict trial effects unique to pragmatic responses to T1 sentences. This resonates with Spotorno and Noveck's (2014) Experiment 3 in which participants, who were exposed to repeated instances of irony-laden sentences (among random filler items), produced *early-late* effects. As the session wore on, participants' readings of ironic sentences sped up. We expect the same outcome here. We expect the cost of making pragmatic interpretations of T1 sentences to decrease across then experimental session, to the point that false responses to T1 should not show signs of requiring added cost, relative to those who repeatedly provide logical readings of T1 sentences, by the end of the session.

Participants

As per our pre-registration, we aimed to recruit at least 175 participants who did not participate in Experiment 1 (but otherwise met the same criteria). This number was based on a power analysis via simulations conducted using the R package SimR (Green and MacLeod 2016). The power analysis involved multiple steps. We first analyzed the 'sweet spot' version of Experiment 1 to test for trial effects (following the procedure described in the following sections). Based on the parameters obtained from the resulting model, we simulated 1000 different data sets extending the number of participants to 175. We then re-fitted the model to each new synthetic data set. Critically, for this step we assumed a more conservative true effect size (for the interaction between STRATEGY and TRIAL NUMBER) half the size of that found in the original analysis. Finally, we counted the number of simulated experiments that showed a significant interaction effect. The

results of this power analysis suggested that power for finding an effect this size or larger should be at least 80% with 175 participants, assuming that the null hypothesis is false. The power analysis is described in detail in the analysis script found on the project's OSF repository.

Anticipating that some of the participants would not meet our exclusion criteria (e.g., if they are not native speakers of English, or if they failed to achieve at least 70% accuracy in the filler items), we recruited a total of 200 people. Of these, none had to be excluded.

Materials, Design & Procedure

Materials, design and procedure were akin to those used for the *sweet spot* version of Experiment 1. The only difference was the additional independent variable TRIAL NUMBER. This represented the order in which a given participant saw each item relative to other items in the same condition, resulting in 9 possible condition numbers for every item (since each participant saw a maximum of 9 items per condition). It is important to elaborate on a critical property of this variable, namely that it is computed relative to other instances of the same condition only. For T1 sentences, this means that the TRIAL NUMBER of T1_pragmatic and T1_logical will be computed independently from one another. We did this because we are interested in the separate cumulative effect of these interpretations, i.e., in the effect that previously deriving a pragmatic interpretation of an underinformative sentence will have on deriving a pragmatic interpretation of a future one. In practice, this means that if, for example, a participant understood the first four T1 sentences logically, but the fifth pragmatically, this fifth T1 would count as the first instance of a T1_pragmatic (out of a maximal possible nine). It also means that those participants who are entirely consistent throughout will reach a TRIAL NUMBER of 9. Otherwise, they will not.

Predictions and Analysis

We had two specific predictions for this study. First, we predicted that we would replicate the results of B&N (Experiment 3) and, of course, those in the *sweet spot* condition of Experiment 1. We expected to find longer reaction times for pragmatic readings of T1 sentences relative to all other conditions. Second, we predicted that this pattern would change as participants encounter new instances of each condition. This prediction is critically based on the results of the *sweet spot* condition of Experiment 1, in which we found a significant interaction between STRATEGY and TRIAL NUMBER. Following up on this finding, we predict that, in Experiment 2, response times to pragmatic readings of T1 sentences should decrease over time relative to a baseline, which we

designated as the response times to the control condition T2. This should result in a significant interaction between TRIAL NUMBER and STRATEGY for the comparison between T1_pragmatic and T2 sentences. Crucially, we predicted no interaction between TRIAL NUMBER and STRATEGY for the comparison between T1_logic and T2 sentences. These pre-registered predictions reflect our expectation that the processing effort typically associated with deriving scalar implicatures can be reduced after repeatedly deriving the same interpretation and at a rate that is above and beyond the shortening of reaction times that could affect the other conditions in this experiment.

To test these predictions, we fitted two linear, mixed-effects models on the log-transformed response times (transformed following the results of a box-cox test, which showed non-normal model residuals and suggested a log-transformation as the optimal transformation of the reaction times). All analyses were conducted using the R programming language (R Core Team, 2020) coupled with R-Studio (RStudio Team, 2020). The following R packages were used: ggplot2 (Wickham, 2016), lme4 (Bates, Sarkar, Bates, & Matrix, 2007), Rmisc (Hope, 2013), MASS (Ripley et al., 2013), dplyr (Wickham, François, Henry, & Müller, 2020), ggpubr (Kassambra, 2020), DoBy (Højsgaard, 2012), papaja (Aust & Barth, 2017), here (Müller, 2017), and afex (Singmann et al, 2020).

Response times were calculated from the point in which the sentence appeared on screen until participants pressed a key. Prior to analysis, we subtracted 1000 milliseconds to all response times (since this was the minimum amount of time participants were required to wait before responding). We then removed all response times longer than 10 seconds and shorter than 200 milliseconds, which amounted to removing 7% of all data points. The first model included STRATEGY (T1-T6) (treatment contrast-coded) and TRIAL NUMBER (centered, continuous predictor) as fixed effects, with random intercepts and slopes by items and by participants for STRATEGY. This model was meant to test our first prediction. To test our second prediction, the model includes STRATEGY (T1_pragmatic, T1_logic and T2, treatment contrast-coded), TRIAL NUMBER (centered, continuous predictor) and their interaction as fixed effects, with random intercepts and slopes for all factors and their interactions by items and by participants (but suppressing the random correlations between intercepts and slopes). This (as well as all other models described in the article) was the 'maximal' model, fitted following the recommendations by Barr, Levy, Scheepers, and Tily (2013). In this model, TRIAL NUMBER was computed independently for T1_pragmatic and T1_logical responders. A consequence of this approach is that not all participants see the same number of T1 conditions (since we have no control over whether

participants interpret a given T1 logically or pragmatically), and only consistent responders will reach TRIAL NUMBER 8 or 9 in the T1_pragmatic and T1_logical conditions. This makes the data from consistent responders particularly important, since for these participants, the comparison between logical, pragmatic and control (T2) sentences is actually balanced. As a post-hoc measure, we also analyzed the consistent responders' responses separately, in addition to our pre-registered analysis.

All models described in this article were fitted using the 'bobyqa' optimizer and increasing the optimizer's maximum number of iterations in order to avoid convergence problems. All models reported in the manuscript converged without errors.

Results

First, descriptive statistics showed that out of a total of 1568 T1 trials, 839 (53.5%) were understood logically and 729 (46.5%) pragmatically. Out of the 91 participants who responded consistently (8 or 9 trials) to T1s, 49 were logical and 42 were pragmatic. This supports the results of Experiment 1, suggesting that a 1-second delay together with a full sentence presentation strikes the *pragmatic sweet spot* for a web-based adaptation of B&N.

The inferential results broadly confirm our predictions. This can be seen in Figure 2. The T1_pragmatic condition elicited significantly longer reaction times relative to all other conditions, replicating the results of B&N (see Table 1). This added cost was significantly reduced as participants derived pragmatic inferences over the course of the experiment. Reaction times to all sentence types generally decreased as the experiment progressed, but this was particularly evident for T1_pragmatic responses, which generally began with the slowest responses overall. This is exemplified by the interaction between TRIAL NUMBER and STRATEGY for the comparison between T1_pragmatic and T2 sentences. No such significant interaction was found between TRIAL NUMBER and STRATEGY for the comparison between T1_logic and T2 sentences (see Table 2). Importantly, this pattern – significant change in processing cost for pragmatic, but not for logical interpretations across TRIAL NUMBER relative to a baseline – held when using any of the other control conditions (T3-T6) as the baseline (all p's<0.05 for the interaction with T1_pragmatic, and all p's>0.05 for interaction with T1_logic, see R-scripts in the online repository).

A post-hoc model also revealed an interaction of TRIAL NUMBER and STRATEGY for the direct comparison between T1_logic and T1_pragmatic sentences (see Table 3). This result

reflects how after repeated exposure, the processing cost of deriving a scalar implicature (relative to understanding the sentence logically) fades. In fact, when taking only the last two trials of the experiment into account, a model with only STRATEGY as a predictor fails to find a significant difference in processing time between logic and pragmatic interpretations of T1 sentences (t=0.1, p=0.874). Finally, analyzing the results of only consistent responders confirmed our findings, as shown in Figure 3 and Tables 4 and 5.

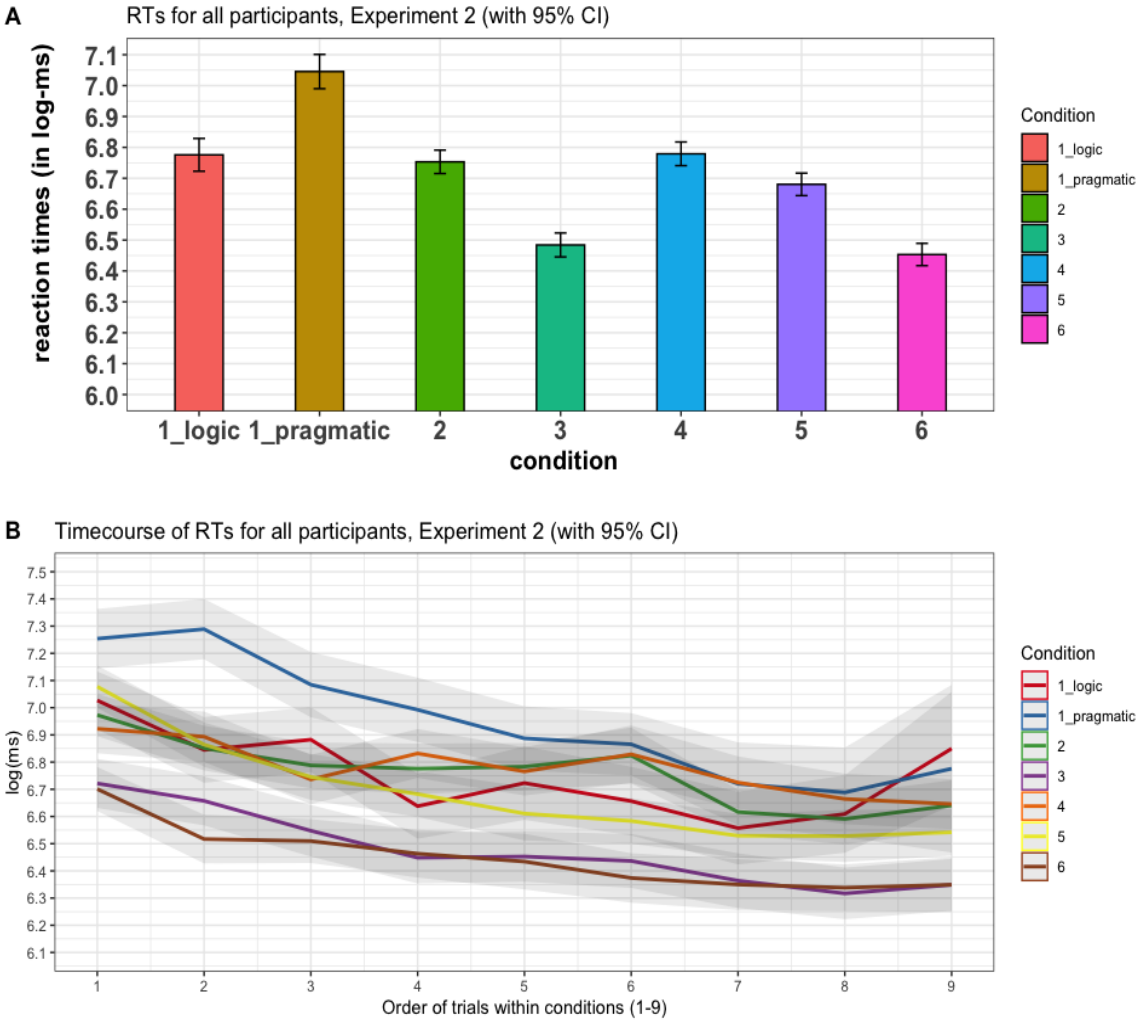

Figure 2. Results of Experiment 2 with all participants without regard to their consistency. 2A refers to overall performance and 2b shows speeds of response over the course of the nine trials that they encounter of each type. Grey ribbons in 2b represent confidence intervals. While there are 9 trials of T1 items, the time courses do not consider individuals, but just the nth instance (1st through 9th) of a trial, without considering whether it prompted a logical or pragmatic response. In principle, a single participant could have provided a true (logical) response at their second encounter with a T1 item and a false (pragmatic) response at their 8th trial.

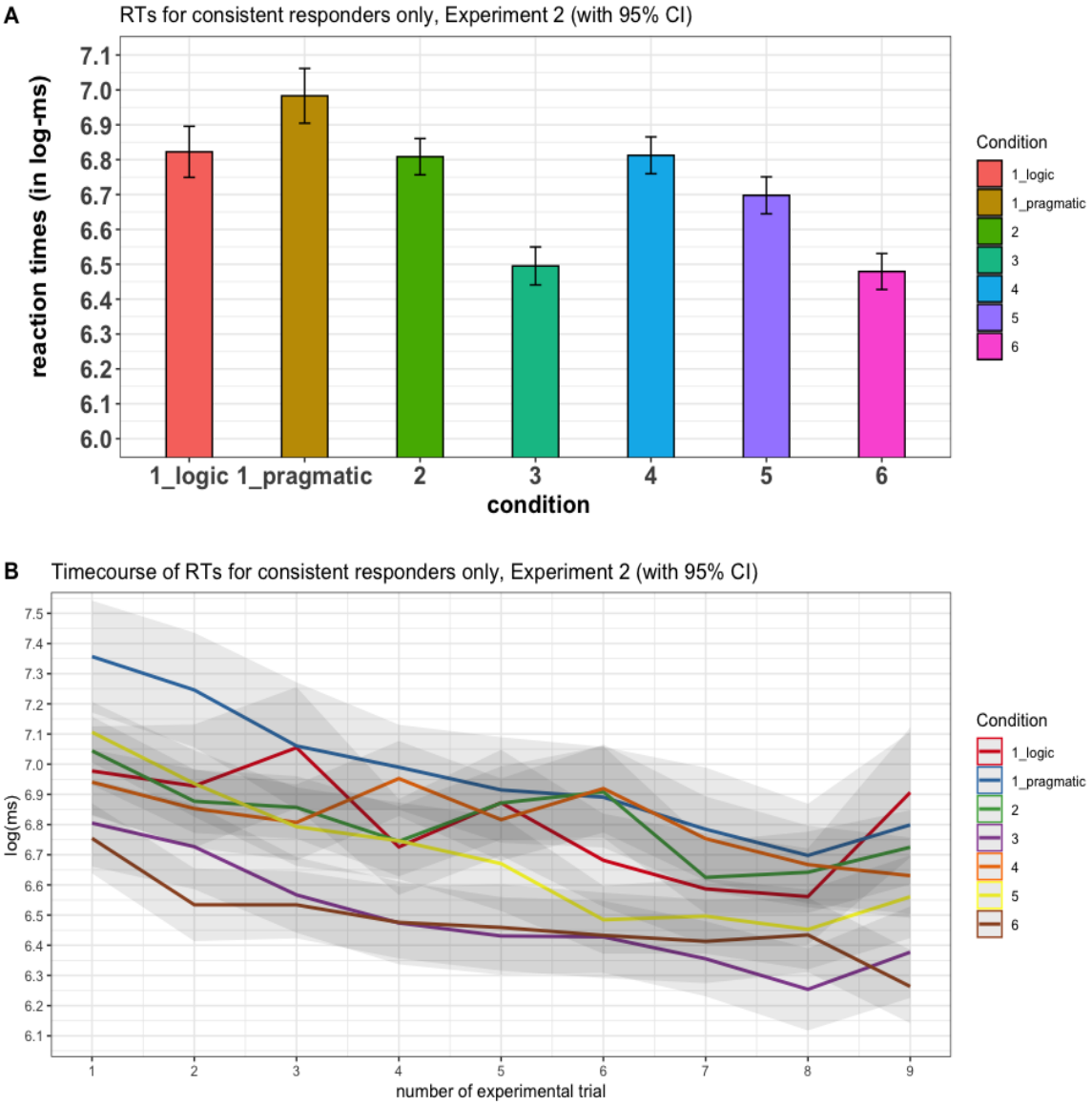

Figure 3. Results of Experiment 2 while considering only consistent responders to the T1 items (those "pragmatic" responders who provide false responses 8 or 9 times out of 9 trials and "logical" responders who provide true responses 8 or 9 times out of 9).

Table 1:

Regression model output replicating the reaction time results of Bott and Noveck (2004)

term	$\hat{\beta}$	95% CI	t	df	p
T1-Pragmatic	7.00	[6.93, 7.06]	209.90	572.19	< .001
… vs. T1-Logic	-0.28	[-0.36, -0.21]	-7.67	176.40	< .001
… vs. T2	-0.23	[-0.29, -0.18]	-7.93	223.88	< .001
… vs. T3	-0.53	[-0.59, -0.46]	-15.46	150.08	< .001
… vs. T4	-0.22	[-0.27, -0.16]	-7.36	265.18	< .001
… vs. T5	-0.31	[-0.37, -0.26]	-11.03	281.79	< .001
… vs. T6	-0.56	[-0.62, -0.50]	-18.49	191.65	< .001
TRIAL NUMBER	-0.12	[-0.13, -0.11]	-19.56	8,442.62	< .001

Note. STRATEGY was treatment-contrast coded. T1-Pragmatic condition coded as the intercept.

Table 2:

Regression model output showing effect of TRIAL NUMBER

Term	$\hat{\beta}$	95% CI	t	df	p
T2	6.77	[6.70, 6.83]	209.47	244.55	< .001
… vs. T1-Logic	-0.02	[-0.09, 0.04]	-0.74	166.75	.459
… vs. T1-Pragmatic	0.18	[0.13, 0.24]	6.36	2,726.81	< .001
TRIAL NUMBER	-0.10	[-0.14, -0.06]	-5.31	121.48	< .001
(T2 vs. T1-Logic)*TRIAL NUMBER	-0.04	[-0.10, 0.02]	-1.20	169.37	.230
(T2 vs. T1-Pragmatic)*TRIAL NUMBER	-0.13	[-0.19, -0.08]	-4.60	84.21	< .001

Note. STRATEGY was treatment-contrast coded. T2 condition coded as the intercept.

Table 3:

Regression model output showing effect of TRIAL NUMBER: second version

term	$\hat{\beta}$	95% CI	t	df	p
T1-Logic	6.75	[6.68, 6.82]	188.56	336.14	< .001
… vs. T2	0.01	[-0.04, 0.07]	0.51	170.64	.614
… vs. T1-Pragmatic	0.21	[0.14, 0.28]	5.69	1,822.19	< .001
TRIAL NUMBER	-0.13	[-0.18, -0.08]	-5.32	299.61	< .001
(T1-Logic vs. T2)*TRIAL NUMBER	0.03	[-0.02, 0.09]	1.20	244.38	.233
(T1-Logic vs. T1-Pragmatic)*TRIAL NUMBER	-0.10	[-0.17, -0.03]	-2.96	159.89	.004

Note. STRATEGY was treatment-contrast coded. T1-Pragmatic condition coded as the intercept.

Table 4:

Regression model output showing effect of TRIAL NUMBER for consistent responders

term	$\hat{\beta}$	95% CI	t	df	p
T2	6.82	[6.73, 6.90]	151.24	122.92	< .001
… vs. T1-Logic	-0.01	[-0.11, 0.09]	-0.20	62.77	.840
… vs. T1-Pragmatic	0.14	[0.06, 0.21]	3.33	110.22	.001
TRIAL NUMBER	-0.10	[-0.15, -0.06]	-4.38	99.84	< .001
(T2 vs. T1-Logic)*TRIAL NUMBER	-0.01	[-0.08, 0.07]	-0.17	89.81	.864
(T2 vs. T1-Pragmatic)*TRIAL NUMBER	-0.12	[-0.19, -0.04]	-3.13	54.83	.003

Note. STRATEGY was treatment-contrast coded. T2 condition coded as the intercept.

Table 5:

Regression model output showing effect of TRIAL NUMBER for consistent responders: second version

term	$\hat{\beta}$	95% CI	t	df	p
T1-Logic	6.82	[6.72, 6.92]	132.38	176.97	< .001
… vs. T2	0.00	[-0.08, 0.08]	-0.08	173.25	.934
… vs. T1-Pragmatic	0.14	[0.03, 0.25]	2.57	202.28	.011
TRIAL NUMBER	-0.11	[-0.17, -0.05]	-3.60	195.34	< .001
(T1-Logic vs. T2)*TRIAL NUMBER	0.01	[-0.06, 0.07]	0.16	1,020.34	.874
(T1-Logic vs. T1-Pragmatic)*TRIAL NUMBER	-0.12	[-0.20, -0.03]	-2.63	846.91	.009

Note. STRATEGY was treatment-contrast coded. T1-Pragmatic condition coded as the intercept.

Discussion

Experiment 2 confirmed our prediction that pragmatic responses to T1 sentences are indeed slower than the other (T2-T6) conditions as B&N reported and, most importantly, they are slower than logical responses to T1. However, a closer look at the progression of pragmatic T1 responses shows that this difference is due to response times in the early part of the task. After multiple encounters with the under-informative T1 sentences, participants' processing times to answer pragmatically progressively attenuates, to the point where after about 7 such sentences, the processing effort appears indistinguishable from that of logical readings of the same sentences. Importantly, this effect plays out differently for logical readings of T1 as well as for the control

sentences: While the processing time to evaluate these sentences also drops with repeated exposure, this reduction is moderate compared to pragmatic readings of T1 sentences. This is in line with our general hypothesis, that pragmatic readings are particularly affected by *early-late effects*, suggesting that as a participant adapts to the intention of the anonymous speaker (which in this case is that the speaker presumably intends the comprehender/participant to understand T1 sentences as false), the processing effort in responding dissipates. This is a critical piece of evidence supporting the claim that slowdowns linked to pragmatic responses to T1 sentences are at least partly due to orienting to a speaker who presumably intends that a false response be derived.

It is important to point out how Figure 2b (all responders) differs from Figure 3b (consistent responders) and what it tells us about the attenuation of pragmatic speeds in this task. Remember that the trial numbers in Figure 2b reflect a participant's encounter with a given type of item. While T2-T6 items are likely to provide consistent responses throughout, the T1 items are not. That is, a participant who at experimental trial 6 is "pragmatic" could be someone who is seeing their ninth and last T1 item and who has also provided three logical responses or it could be a perfectly consistent pragmatic responder's sixth encounter. In contrast, Figure 3b shows only consistent responders. Note that both Figures are revealing of an early-late effect. When the inconsistent responders are included (in Figure 2b) the convergence arrives later.

In light of these data, it is hard to argue that slowdowns are due solely to participants mechanically enriching *Some* to mean *not all*. In the next Experiment, we examine in fine detail whether this *early-late effect* can be made contingent on the presentation of specific speakers.

Experiment 3: Speaker-specific pragmatic processing costs

Experiment 2 aimed to establish that the processing cost typically associated with deriving a scalar implicature in the B&N task fades as the experiment progresses. As outlined in the Introduction, we believe that these *early-late effects* are caused by how participants engage in mindreading: After repeatedly understanding T1 sentences pragmatically, the cost of grasping the speaker's informative intention dissipates, resulting in similar processing costs for logical and pragmatic interpretations. To provide further evidence that these pragmatic effects are due to determining the speaker's informative intention, Experiment 3 tests a critical consequence of our hypothesis: If the cost of interpreting a T1 sentence pragmatically is related to reading a speaker's intentions, the introduction of a new speaker mid-way should counteract any *early-late effects* and re-set the processing cost. In this pre-registered study, we therefore compare two conditions. In one, there is a single "speaker" who is simply presented at the end of the training phase. In the second condition, there are two speakers, with a second speaker introduced midway in the task. Our prediction is that the introduction of a second speaker will reboot the intention-reading processes and thus recreate a full-blown T1-related slowdown, which should not occur when the speaker remains one and the same.

Participants

Following our pre-registration, we recruited 500 participants who did not participate in Experiments 1 or 2. This number was based on a power analysis via simulations performed on pilot data, in a fashion similar to what was done for Experiment 2. Data from the pilot study and R-script for the power analysis can be found in the project's OSF repository.

Materials, Design & Procedure

Materials, design and procedure were similar to those of Experiment 2. There were three main differences. First, we introduced a between-subjects manipulation, the factor SPEAKER (levels: "One Speaker" vs. "Two Speakers"). After going through the same practice round as in Experiment 2, participants in the "One Speaker" condition (N=250) were presented with a cover story introducing a specific person said to be the producer of all the statements. The speaker presented was one of two possibilities: A young man described as an athlete and sports fan or a grandmother who enjoys knitting (cover stories can be found in the OSF repository). The cover

stories were counterbalanced so that 125 participants saw the 'athlete' and 125 the 'granny' cover story. Participants in the 'Two Speakers' condition (N=250) were told that they would read statements produced by two different people. At the beginning of the Experiment they were shown one of the cover stories ('granny' or 'athlete') and the other one half-way through the experimental session. Order of presentation of cover stories was counterbalanced.

The second difference was the block structure and pseudo-randomization scheme. Experiment 3 only had two blocks instead of the three used in Experiments 1 and 2 (one block per potential speaker in the two-speaker condition). The pseudo-randomization scheme was changed so that participants would always see five T1 (and five T2) sentences in the first block and the remaining four in the second.

Finally, the predictor TRIAL NUMBER was nested within blocks to better reflect the structure of the experiment: Instead of counting how many logical or pragmatic readings of T1s occurred from 1-9, we counted how many occurred in Block 1 (1-5) and Block 2 (6-9) separately. For example, if a participant understood the first five T1s logically and the sixth pragmatically, this sixth T1 would not be assigned the number 1 (as would have been in Experiment 2), but the number 6 (i.e., representing the first T1_pragmatic instance of Block 2). This was done because we were interested in the potential change in processing cost when switching blocks (i.e. speakers), which happened when going from TRIAL NUMBER 5 to 6, and how this compares to differences in switching blocks when the speaker is the same (changes in TRIAL NUMBER 5 to 6 in the 'Single Speaker' condition). Using the same TRIAL NUMBER measure as in Experiment 2 would have made it difficult to observe potential change, since we could not have known for certain if a specific T1_pragmatic came before or after the change in speakers, blurring the line that separates Blocks 1 and 2.

Predictions and Analysis

We analyzed the data according to our main hypothesis, namely that there should be a difference in the way that trial effects for pragmatic interpretations of T1 sentences develop between a single-speaker and a two-speaker set-up. To test this, we examined T1 and T2 control sentences. We predicted that the response times to T1_pragmatic (relative to the T2 control condition and to the T1_logical condition) would show a significant three-way interaction with TRIAL NUMBER and SPEAKER. This prediction is based on the results of our pilot Experiment (N=100), for which this three-way interaction was found (see analyses on the OSF repository). In this interaction, we

predict that the difference between T1_pragmatic and T2 controls should shorten as the participant receives more of such sentences in the 'single speaker' condition but not in the 'two-speaker' condition, considering how in the 'two-speaker' condition, the processing cost should increase again when the participant arrives at TRIAL NUMBER 6 (i.e., at the beginning of the second block). In other words, the presence of the second speaker arriving midway should re-start the participant's efforts to understand the speaker's intention. No such three-way interaction effect should be present between T1_logical (relative to the T2 control condition), TRIAL NUMBER, and SPEAKER.

We first removed trials with reaction times longer than 10s seconds and shorter than 200 milliseconds, leading to the exclusion of 7.2% of the data. We then fitted a mixed-effects, linear regression model to the data. The model used a treatment contrast coding scheme for the factor STRATEGY and a sum-contrast coding scheme for the factor SPEAKER. The continuous predictor TRIAL NUMBER was centered. The model was fitted with the level 'T1_pragmatic' of the factor STRATEGY as the baseline and had a 'maximal' random effects structure (minus the random correlations between intercepts and slopes). As was done for Experiments 1 and 2, we log-transformed the response times (given non-normality of the residuals).

We also conducted two additional post-hoc tests to probe the critical moment where participants switch from Block 1 to Block 2. First, we created a subset of the data to keep only the response times of the last T1_pragmatic, T1_logical and T2 in Block 1 and the respective first ones of Block 2. We then fitted a model to this subset with BLOCK (1 vs. 2), SPEAKER ('single speaker' vs. 'two speakers'), STRATEGY ('T2' vs. 'logical' vs. 'pragmatic') and their interactions as predictors. BLOCK and SPEAKER were sum-contrast coded. STRATEGY was helmert-contrast coded. Helmert contrast compares the second level of a categorical predictor with the first, and the third with the average value of the first two. This allows us to compare 'T2' and 'T1_logical' to each other (coded as the first and second variable) and 'T1_pragmatic' (coded as the third) to the average value of the other two. The prediction here is the following: If switching blocks in the 'two speakers' case is particularly costly for T1_pragmatic sentences but not for T1_logical or T2 controls, then there should be a three-way interaction between BLOCK, SPEAKER, and STRATEGY ('T2' and 'logical' vs. 'pragmatic'), but not a three-way interaction between BLOCK, SPEAKER, and STRATEGY ('T2' vs. 'logical'). The random effects structure here was 'maximal', with the exception of excluded random correlations between intercepts and slopes.

Our second post-hoc test was identical to the previously described one but focused on the consistent responders. For consistent responders, the 5th and 6th instance of T1_pragmatic and T1_logic coincide with the last T1 sentence from the first block and the first T1 sentence of the second block. Changes in processing effort between 5th and 6th TRIAL NUMBER for this group of responders would therefore more directly reflect the cost of changing blocks, since they represent adjacent T1 trials for these participants.

Results

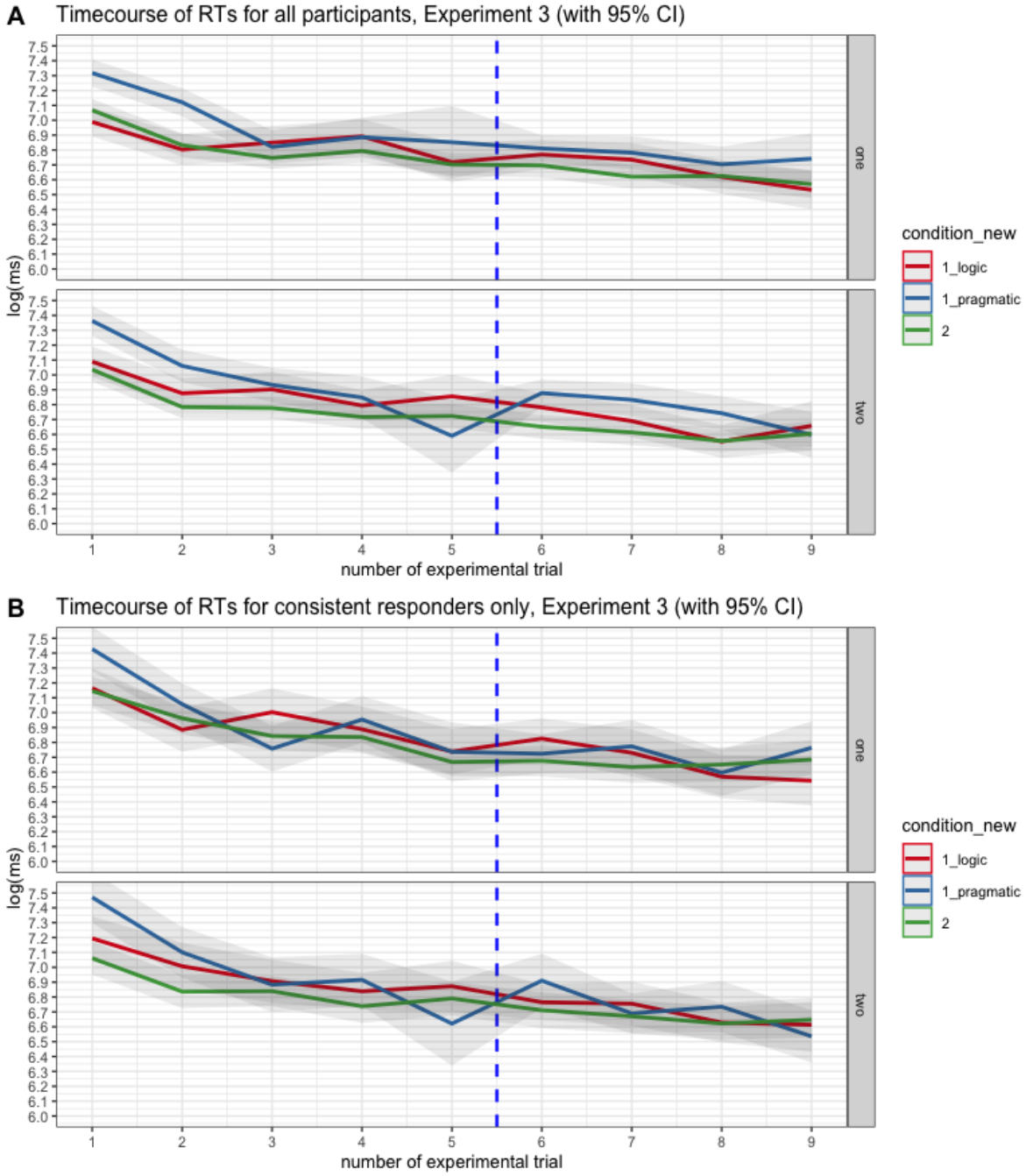

Figure 4. Results of Experiment 3. Dotted line represents the end of Block 1.

Table 6:

Experiment 3: Regression model output, all participants

term	$\hat{\beta}$	95% CI	t	df	p
T1-Prag. v. T2	-0.23	[-0.28, -0.18]	-9.34	140.91	< .001

term	$\hat{\beta}$	95% CI	t	df	p
T1-Prag. v. T1-logic	-0.21	[-0.26, -0.17]	-8.87	1,117.18	< .001
TRIAL NUM.	-0.19	[-0.22, -0.16]	-12.77	1,888.44	< .001
SPEAKER	-0.04	[-0.12, 0.05]	-0.82	1,082.14	.413
(T1-Prag. v. T2) X TRIAL NUM.	0.06	[0.03, 0.10]	3.59	6,518.45	< .001
(T1-Prag. v. logic) X TRIAL NUM.	0.05	[0.01, 0.10]	2.45	221.37	.015
(T1-Prag. v. T2)*SPEAKER	0.05	[-0.03, 0.13]	1.22	1,001.61	.224
(T1-Prag. v. logic) X SPEAKER	0.00	[-0.09, 0.10]	0.07	1,111.34	.945
TRIAL NUM. X SPEAKER	0.00	[-0.06, 0.06]	-0.04	431.34	.969
(T1-Prag. v. T2) X SPEAKER X TRIAL NUM.	-0.01	[-0.08, 0.06]	-0.29	248.17	.768
(T1-Prag. v. logic) X SPEAKER X TRIAL NUM.	0.04	[-0.04, 0.12]	0.96	1,143.95	.340

Note. STRATEGY was treatment-contrast coded. T1-pragmatic condition coded as the intercept.

Table 7:

Experiment 3: Regression model output showing only Trials 5 and 6, all participants.

term	$\hat{\beta}$	95% CI	t	df	p
BLOCK	-0.01	[-0.09, 0.06]	-0.31	615.11	.757
SPEAKER	0.00	[-0.10, 0.10]	0.03	728.63	.974
T2 vs. Log.	0.02	[-0.03, 0.06]	0.79	50.42	.431
T2+Log. vs. Prag.	0.05	[0.02, 0.08]	2.89	119.44	.005
BLOCK*SPEAKER	-0.05	[-0.20, 0.10]	-0.64	620.15	.524
BLOCK*(T2 vs. Logical)	-0.02	[-0.09, 0.05]	-0.67	488.33	.503
BLOCK*(T2+Log. vs. Prag.)	-0.02	[-0.08, 0.05]	-0.58	98.88	.562
SPEAKER*(T2 vs. Logical)	0.04	[-0.05, 0.13]	0.97	71.72	.335
SPEAKER*(T2+Log. vs. Prag.)	-0.05	[-0.12, 0.02]	-1.48	601.03	.139
BLOCK*SPEAKER*(T2 vs. Logical)	0.04	[-0.13, 0.21]	0.50	56.76	.618
BLOCK*SPEAKER*(T2+Log. vs. Prag.)	-0.16	[-0.28, -0.03]	-2.45	99.30	.016

Note. BLOCK and SPEAKER were sum-contrast coded. STRATEGY was helmert-contrast coded.

Table 8:

Experiment 3: Regression model output showing only Trials 5 and 6, consistent readers only.

term	$\hat{\beta}$	95% CI	t	df	p
BLOCK	-0.03	[-0.12, 0.05]	-0.77	304.32	.443
SPEAKER	0.10	[-0.04, 0.23]	1.43	284.92	.155
T2 vs. Log.	0.02	[-0.02, 0.07]	1.03	192.00	.303
T2+Log. vs. Prag.	0.03	[-0.01, 0.07]	1.52	490.39	.130
BLOCK*SPEAKER	-0.02	[-0.20, 0.16]	-0.25	63.02	.806
BLOCK*(T2 vs. Logical)	-0.01	[-0.09, 0.08]	-0.13	369.72	.900
BLOCK*(T2+Log. vs. Prag.)	-0.05	[-0.12, 0.02]	-1.39	73.64	.167
SPEAKER*(T2 vs. Logical)	-0.03	[-0.14, 0.09]	-0.46	43.20	.647
SPEAKER*(T2+Log. vs. Prag.)	-0.02	[-0.09, 0.06]	-0.44	496.29	.659
BLOCK*SPEAKER*(T2 vs. Logical)	0.10	[-0.11, 0.31]	0.91	39.01	.368
BLOCK*SPEAKER*(T2+Log. vs. Prag.)	-0.17	[-0.32, -0.03]	-2.35	47.77	.023

Note. BLOCK and SPEAKER were sum-contrast coded. STRATEGY was helmert-contrast coded.

Figure 4 illustrates the results of Experiment 3. The statistical analyses are reported in Tables 6-8. As Table 6 shows, we replicated the B&N effect: pragmatic responses were significantly delayed relative to T2 controls (t-value=9.34, p<0.001) and T1_logical sentences (t-value=8.87, p<0.001). We also found an interaction between TRIAL NUMBER and STRATEGY (when comparing T1_pragmatic vs. T2) (t-value=3.6, p<0.001) and an interaction between TRIAL NUMBER and STRATEGY (when comparing T1_pragmatic vs. T1_logic) (t-value=2.45, p<0.01). We did not find the predicted significant three-way interaction between STRATEGY (T1_pragmatic vs. T1_logic, or T2 vs. T1_pragmatic), SPEAKER and TRIAL NUMBER (t-value =0.96, p=0.34).

Despite the absence of this three-way interaction, visual inspection of Figure 4 suggests that the processing effort of T1_pragmatic sentences (relative to that of T1_logical) changes in the 'two speakers' condition when going from Block 1 to Block 2. Apparently, our registered statistical prediction was too strong (it considered the entire length of the experiment as opposed to focusing on the moment when the speakers were switched) for detecting an effect that is localized to the moment in which participants switch between Blocks 1 and 2. When we focus on the transition point (between the 5th and 6th trials), we find support for our claims.

We ran two tests, which are summarized in Tables 7 and 8. The first model, reported in Table 7, showed a significant three-way interaction (t-value=2.45, p=0.016) between TRIAL NUMBER, STRATEGY (T2+logical vs. pragmatic) and BLOCK, but no significant interaction between

TRIAL NUMBER, STRATEGY (T2 vs. logical) and BLOCK (t-value=0.5, p=0.6). The second model, reported in Table 8, duplicates this analysis but considers consistent speakers only, with identical results. This is in line with what can be seen in Figure 4, namely, that the relationship between pragmatic readings, on the one hand, and both logical reading and control sentences (but not between logical readings and control sentences), on the other, changes across blocks, but only for the 'two speakers' case.

At this point, it could be argued that the significant three-way interaction seen in Table 7 is driven simply by unusually fast responses in the last trial of Block 1 and not by the switch from Block 1 to Block 2. To rule out this possibility, we conducted a further test to assess whether the transition from the fourth to the fifth trial was noticeably steeper in the 'two speaker' relative to the 'single speaker' condition. We fitted the identical model summarized in Table 7 but instead compared trials 4 and 5 (instead of trials 5 and 6). This analysis did not find a significant three-way interaction (p=0.085). This is consistent with our claim that it is only upon changing blocks that the processing cost associated with T1_pragmatic sentences is 'renewed'.

Discussion

Experiment 3 was concerned with the change in processing effort of pragmatic interpretations of underinformative sentences in relation to the number of speakers believed to be the producers of such sentences. We hypothesized that if there is only a single producer throughout an experimental session, the processing cost of pragmatic, relative to logical, readings of underinformative sentences (as well as to control sentences) should decrease, as was the case in Experiment 2. To start, we did confirm the same interaction found in Experiment 2: Overall, the processing cost of pragmatic readings of T1 sentences decreases across the experimental session relative to both the logical readings of the same sentences and the control sentences.

To assess the influence of the change of speakers, we conducted two types of tests. The first was based on a severe measure that predicted a significant three-way interaction when the analysis takes into account all trials. That we would find such strong effects appears to have been overly ambitious. Our second type of test considered just the critical moment between a participant's 5^{th} and 6^{th} encounter with the T1 sentences. In the two-speaker condition, this marked the moment when a speaker was switched; in the meanwhile, nothing changed for the participant in the one-speaker condition.

This analysis supports our hypothesis, suggesting that upon switching speakers, the processing cost for pragmatic utterances appears to rebound, resulting in processing time differences between the last T1 trial of the first block and the first T1 trial of the second. This is consistent with our claim that at least part of the processing cost typically associated with deriving scalar implicatures can be attributed to the moment participants reason about the intentions of individual speakers that lie behind the production of underinformative sentences.

General Discussion

Since Bott & Noveck (2004) researchers have persistently investigated whether pragmatic readings of underinformative sentences prompt slowdowns compared to controls. While there is much evidence in support of a claim that defends effortful pragmatic processing generally, even as tasks became more plentiful and varied, there is also research that has chipped away at this claim (Degen & Tanenhaus, 2015; Grodner et al., 2010). In the present work, we argue that the pragmatic delay in B&N task, which has remained central to such claims, actually brings mindreading into play and that it accounts for much of the processing cost as scalar implicatures are derived.

In describing our account, we pointed out how the anonymity and the apparent equivocality of the hidden speaker in the B&N task makes it difficult for participants to readily attribute intentions. The only piece of evidence that participants have at their disposal is that the speaker presents statements that are true or else false in roughly equal measure (by the time participants receive their first T1 sentence they have seen both kinds). Participants are thus in a position to consider two strategies when interpreting T1 sentences, one that says that the speaker is inclined to provide statements that are true and another that says that the speaker is inclined to provide statements that are false. If a participant adopts the latter, they most likely will adjust their reading of the underinformative sentence in T1 so that it is consistent with the false response. So, while pragmatic responses entail making a modification to what is said, it is also the case that pragmatic responders are making attributions about the speaker's intention.

We carried out three experiments. The first was practical in that we aimed to come up with a version of the B&N task that reliably provides the kind of data found in laboratory versions, i.e. roughly equal rates of pragmatic and logical responses as well as slowdowns related to the former. This led us to adopt a version that presents each item as a single sentence along

with a one-second delay between the presentation of the sentence and the availability of response options. At this point, we began to explore our hypothesis.

Experiment 2 proposed and tested the idea that – given the mindreading aspect of the task – one should find *early-late effects* that are unique to pragmatic responses. Indeed, we found that pragmatic responses to T1 sentences speed up over time at a higher rate than logical interpretations of the same sentences and that all other control sentences. This is in itself a valuable finding for the literature because it modifies our understanding concerning slowdowns. Notably, this effect was maintained in Experiment 3 when we literally introduced a face and description, which brought the speaker out of the shadows.

For Experiment 3, we hypothesized that if consideration of a speaker's intention plays a role in the derivation and slowdown of a pragmatic response to T1 items, then the mere introduction of a new speaker should prompt participants to renew their search for a speaker's intention and, again, add a cost. To test this hypothesis, we prepared two conditions. In one condition, we presented one plainly described person (along with a photo) at the start of the session. In the other, we provided two plain descriptions (along with a photo for each) that were introduced at two key points: one at the outset of the task and another at a precise point midway (after the fifth encounter with a T1 sentence). We found that there was an uptick in T1 pragmatic response latency on the sixth encounter (i.e., after the new speaker is introduced) and for the two-speaker condition only.

What exactly is slowing pragmatic responders down? Based on the current data, it is hard to argue that structural features that are related to the generation of a scalar implicature are its only source of processing effort, as is commonly argued in the literature. For example, Bott et al. (2012), following up on Bott & Noveck (2004), proposed that slowdowns are related to the fact that pragmatic participants realize that their reading of *some* ultimately leads to an empty set. Accordingly, a pragmatic reading of *Some cats are tabby* considers tabby cats (reference set) and those cats that are not tabby (complement set); with *Some cats are mammals*, the complement set - cats that are not mammals – prompts a false response because, empirically, there are no such cases. For this step to be considered integral to slowdowns, it would presumably occur each time a pragmatic reading is reached, given that the participant must compare the complement set to an empty one. This is incompatible with our current results.

It is also telling that the long pragmatic latencies reported here appear short-lived in the context of an experimental session. We argue that the long initial pragmatic response times to T1

items reflect participants' (addressees') attempts to adjust to the speaker's informative intention: Slowdowns are (at least partly and potentially entirely) due to modifying the reading of the T1 statement (to provide a pragmatic enrichment) *as* a participant works from the speaker's supposed informative intention (that the T1 item is false). In other words, we argue that pragmatic adjustments are *in the service* of what the participant considers to be the informative intention. The *early-late effects* show that, once the pragmatic responder identifies what they consider to be the speaker's informative intention early in the task, responding false to subsequent T1 items ought not to be as time consuming.

How does one characterize the speaker-specific adjustment? That is, what is the starting point that would make a participant's *initial reaction* to a speaker presenting a T1 item appear so effortful? We have two hypotheses. One is that addressees' initial stance assumes that incoming information is truthful and that *any* false reading requires effort (in line with Grice's Cooperative Principle). However, as has been pointed out elsewhere (Bott & Noveck, 2004), if this were the case one would not find such fast rejections to the T3 and T6 sentences which are false on their face. It follows that the *nature* of the rejected T1 sentence plays a role in response times in the B&N task. This leads to our second hypothesis, which is that a TRUE response to a T1 item must appear prepotent to a pragmatic responder and that this, at least initially, prompts interference for FALSE responders. We consider the relatively fast logical T1 responses, which is another feature of B&N studies that remains prominent in our online studies, as evidence indicating that a logical reading interferes with those who ultimately provide the slow false response. This is in line with two-step accounts of the B&N effects (Tomlinson et al., 2013), according to which participants are initially attracted to a true reading before taking into account a pragmatic reading. Based on our approach, such effects ought to be short lived.

While the current findings and claims concerning the B&N task are original for the scalar implicature processing literature, there are in fact other observations in the literature that resonate with ours. First off, Fairchild and Papafragou (2021) found that the derivation of scalar implicatures (using B&N-style materials) correlated positively with a measure of Theory of Mind abilities. This supports the idea that pragmatic interpretations of T1 sentences critically involve reasoning about the intentions of a speaker. Further, consider the findings from Grodner et al.'s (2010) eye-tracking study (which required participants to follow instructions such as "click on the girl who has summa the balls" when a competitor girl had all the *ball*oons and a third had no items). They reported (page 50) that "the average target proportion in the quantifier region was

higher in the first half compared to the second half of the experiment." At the time, they presented these findings to argue against a claim that said participants *build up* a strategy to adopt "summa" to mean essentially "only some." Another way to view Grodner et al.'s data is to consider them in line with our account, as indicating that participants work out the speaker's intended meaning (concerning which girl to click on in conjunction with what is said) early in the task as they aim to better decipher the critical word *summa*. Given that the eye-tracking task's statements are not viewed as potentially equivocal and that one still finds *early-late* effects can indicate that such effects might be present with other non-B&N type tasks (including perhaps on vignette-reading tasks).

Consider, too, Breheny et al. (2013), who showed that an addressee's awareness of a speaker's epistemic state is integral to the addressee's implicature processing. The authors showed a video clip portraying the unfolding of a scene (about two boxes and cutlery) to two people, each on their own screen (e.g. they see an actor's hand place a fork in Box A, then a fork in Box B, and then a spoon in Box A and in that order). However, only one of the two people sees the scene through to the end; the other person -- a confederate who will later be the speaker -- has her view ostensibly blocked after the two forks are placed in their respective boxes and thus does not see anything concerning the spoon (importantly, both the confederate-speaker and the participant see that the former is blocked from seeing the end of the video). When the speaker says "There is a fork in box …." the addressee does not anticipate looking at Box B (the fork-alone box) as is their wont when both viewers see the scene to completion. That is, participants' anticipatory looks are dependent on recognizing the speaker's current informative intention.

For the moment, it is hard to know the extent to which speaker-specific effects, e.g., in the form of *early-late* effects, are present generally across scalar implicature studies. The current study benefited from being online and having hundreds of participants, giving it enough power to make trial effects evident. As online studies become more common, it will be in researchers' interest to pay closer attention to this effect, especially for tasks that have the sentence-verification features of the B&N task. As we reported, trial effects have been found in other pragmatic tasks, such as irony (Spotorno & Noveck, 2014; see also Olkoniemi et al., 2016). Future work will determine the extent to which pragmatic tasks prompt *early-late* effects generally and under what conditions.

Before we conclude, we would like to address what some might see as a potential shortcoming of our experiments as well as alternative interpretations to our findings, as generated

by helpful reviews. One issue worth addressing concerns the forced response delay of one second, which was adopted as we sought the pragmatic sweet spot of Experiment 1 and applied it to Experiments 2 and 3. According to one reviewer, the one-second delay might have obscured what would have been particularly fast logical responses to T1 sentences towards the end of the experiment and thus have influenced our pattern of results. We provide three reasons why we think that this is unlikely. First, B&N's Experiment 3 (the basis for the current investigation) reported that RTs for T1 logical responses were on average 2600 ms. This prompted us to assume that, when the sentence is presented as a whole, a one-second delay would not have a noticeable impact on the fastest participants, since participants would likely need at least one second to read and understand the sentence. Second, this was confirmed by our Experiment 2, where the lower-bound of the confidence interval for T1's logical responses was at 1100 milliseconds after the end of the forced pause, meaning that participants appear to require at least 2100 milliseconds after sentence presentation to make a decision (the lower bound CI in Experiment 3 was 40 milliseconds slower). Finally, in Experiments 2 and 3, logical interpretations of T1 sentences were always significantly slower than the T3 and T6 sentences throughout the experiment. This makes it very unlikely that our forced pause obscured the fastest responses: If this had been the case, fast responses for T3 and T6 sentences should have also been obscured by the forced pause, and all three sentence types should be producing similar 'floor' effects.

Here we turn to a potential alternative explanation of our data, which is that it could be the case that it is not mindreading but some other factor – perhaps one typically claimed to be crucial to scalar implicature processing – that is behind our early-late effects. Once this general possibility is taken on board, one could point to the *retrieval of alternatives* to the scalar term (e.g., Bott and Frisson, 2022) or to the *suppression of the literal meaning* (Tomlinson et al., 2013) as being potentially responsible for the early-late effects. The argument goes that each of these processes could in theory get easier with time and reduce the cost of a pragmatic interpretation. Let us consider each in turn.

Regarding the retrieval of alternatives, it has been suggested that processing a scalar term such as *some* typically activates the stronger term *all* (see, e.g., De Carvalho et al., 2016; Ronai and Xiang, 2023). Note that if this were the case – and if retrieving alternatives were to get easier with time -- this should also occur for T2 (*Some mammals are cats*) and T3 sentences (*Some cats are insects*), since both of these items should also activate the stronger alternative *all*. While RTs of these items do get shorter over time, the decrease is steeper for T1 pragmatic sentences

(judging by the interaction of both T2 and T3 sentences vs. T1_pragmatic and trial order). This suggests that the ease of retrieving alternatives is not a sufficient explanation for the size of the critical trial effects in our Experiments.

Regarding suppression, it is an open question as to whether this process should get easier with task-devoted time. Take, for example, the Stroop effect, which arguably involves a prototypical case of suppression. Here, it has been found that, under certain circumstances (when the to-be-ignored word in one trial is the to-be-named color in the next), the effect size actually *increases* from trial to trial, in what Effler (1977) calls 'serial interference' (see also Neill, 1977; Neill & Westberry, 1987; and MacLeod, 1991, for a review). So it is not immediately clear that suppression effects should ease with time devoted to a given task.

While we remain skeptical that traditional mechanically-inspired accounts of scalar-related slowdowns can readily address the early-late effects we report here, we do look forward to future studies that can carefully compare accounts. One question to consider for such future work concerns the extent to which the early slowdowns are maintained under B&N's task conditions. While pragmatic responses prompt initial slowdowns, it also appears --- in some cases at least -- that the slowdowns persist across trials (e.g. see Figure 4A). As one reviewer pointed out, such outcomes could arguably be taken to show that slowdowns are indeed linked to mechanical steps.

Conclusion

The underinformative (T1) sentences examined by Bott and Noveck (2004) have been crucial to advancing discussions in the linguistic-pragmatic literature, since they provide researchers with a prime example of the effortful pragmatic processing that arises when participants interpret these sentences as *false*. While the derivation of the scalar implicature itself is arguably a part of participants' processing in these cases, it is not clear that it counts for all, or even most, of the extra effort they generate. Once the critical (T1) items in the B&N task are viewed with a wider Gricean lens, i.e. by considering not only the presented sentence but the presumed intention of a speaker who utters it, one is in a better position to appreciate the potential sources of the well-documented slowdowns. Here, we showed that these slowdowns arise early on in an experimental session (where they appear to occur intensively and briefly) and arguably because participants are aligning with what they consider to be a speaker's informative intention.

References

Aust, F., & Barth, M. (2017). *Papaja: Prepare reproducible APA journal articles with R Markdown. R package version 0.1. 0.9997*.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Sarkar, D., Bates, M. D., & Matrix, L. (2007). The Lme4 package. *R Package Version*, *2*(1), 74.

Bott, L., Bailey, T. M., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language*, *66*(1), 123–142. https://doi.org/10.1016/j.jml.2011.09.005

Bott, L., & Frisson, S. (2022). Salient alternatives facilitate implicatures. *Plos one*, *17*(3), e0265781.

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

Breheny, R. (2006). Communication and folk psychology. *Mind & Language*, *21*(1), 74–107.

Breheny, R., Ferguson, H. J., & Katsos, N. (2013a). Taking the epistemic step: Toward a model of on-line access to conversational implicatures. *Cognition*, *126*(3), 423–440. https://doi.org/10.1016/j.cognition.2012.11.012

Breheny, R., Ferguson, H. J., & Katsos, N. (2013b). Investigating the timecourse of accessing conversational implicatures during incremental sentence interpretation. *Language and Cognitive Processes*, *28*(4), 443-467.

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, *100*(3), 434–463.

Brown-Schmidt, S. (2009). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language*, *61*(2), 171–190.

De Carvalho, A., Reboul, A. C., Van der Henst, J. B., Cheylus, A., & Nazir, T. (2016). Scalar implicatures: The psychological reality of scales. *Frontiers in psychology*, *7*, 1500.

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology*, *54*(2), 128.

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive Science*, *39*(4), 667–710.

Dieussaert, K., Verkerk, S., Gillard, E., & Schaeken, W. (2011). Some effort for some: Further evidence that scalar implicatures are effortful. *Quarterly Journal of Experimental Psychology*, *64*(12), 2352–2367.

Drummond, A. (2013). Ibex farm. *Online Server: Http://Spellout. Net/Ibexfarm*.

Effler, M. (1977). Experimental contributions toward an analysis of the interference phenomenon observed with the Stroop Test. *Zeitschrift fuer Experimentelle und Angewandte Psychologie*. 24, 244-281.

Fairchild, S., & Papafragou, A. (2021). The role of executive function and theory of mind in pragmatic computations. *Cognitive Science*, *45*(2), e12938.

Filik, R., & Moxey, L. M. (2010). The on-line processing of written irony. *Cognition*, *116*(3), 421–436.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron*, *50*(4), 531–534.

Gardner, B., Dix, S., Lawrence, R., Morgan, C., Sullivan, A., & Kurumada, C. (2021). Online pragmatic interpretations of scalar adjectives are affected by perceived speaker reliability. *PloS One*, *16*(2), e0245130.

Giora, R., & Fein, O. (1999). Irony: Context and salience. *Metaphor and Symbol*, *14*(4), 241–257.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Grice, H. P. (1989). *Studies in the Way of Words*. Cambridge: Harvard University Press.

Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, *116*(1), 42–55.

Grodner, D., & Sedivy, J. C. (2011). The Effect of Speaker-Specific Information on Pragmatic Inferences. In Gibson & Pearlmutter (Eds.), *The Processing and Acquisition of Reference* (pp. 239–272). MIT Press. https://doi.org/10.7551/mitpress/9780262015127.003.0010

Højsgaard, S. (2012). The doBy package. *R Package Version*, *4*(3).

Heyman, T., & Schaeken, W. (2015). Some differences in some: Examining variability in the interpretation of scalars using latent class analysis. *Psychologica Belgica*, *55*(1), 1.

Hope, R. M. (2013). Rmisc: Ryan miscellaneous. *R Package Version*, *1*(5).

Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics/pragmatics interface. *Cognitive Psychology*, *58*(3), 376–415.

Huang, Y. T., & Snedeker, J. (2018). Some inferences still take time: Prosody, predictability, and the speed of scalar implicatures. *Cognitive Psychology*, *102*, 105–126.

Kassambara, A. (2020). ggpubr:"ggplot2" based publication ready plots. R package version 0.4. 0, 438.

Khorsheed, A., Rashid, S. M., Nimehchisalem, V., Imm, L. G., Price, J., & Ronderos, C. R. (2022). What second-language speakers can tell us about pragmatic processing. *PLOS ONE*, *17*(2), e0263724. https://doi.org/10.1371/journal.pone.0263724

Khorsheed, A., Price, J. & Van Tiel, B. (2022). Sources of Cognitive Cost in Scalar Implicature Processing: A Review. *Frontiers Communication*.

Kronmüller, E., & Barr, D. J. (2015). Referential precedents in spoken language comprehension: A review and meta-analysis. *Journal of Memory and Language*, *83*, 1–19.

Kronmüller, E., & Noveck, I. (2019). How do addressees exploit conventionalizations? From a negative reference to an ad hoc implicature. *Frontiers in Psychology*, *10*, 1461.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological bulletin*, *109(2)*, 163.

Marty, P. P., & Chemla, E. (2013). Scalar implicatures: Working memory and a comparison with only. *Frontiers in Psychology*, *4*, 403.

Müller, K. (2017). *Here: A simpler way to find your files* [Manual].

Neill, W. T. (1977). Inhibitory and facilitatory processes in selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(3), 444.

Neill, W. T., & Westberry, R. L. (1987). Selective attention and the suppression of cognitive noise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(2), 327.

Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds.* Clarendon Press/Oxford University Press.

Noveck, I. (2018). *Experimental pragmatics: The making of a cognitive science*. Cambridge University Press.

Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language*, *85*(2), 203–210.

Olkoniemi, H., Ranta, H., & Kaakinen, J. K. (2016). Individual differences in the processing of written sarcasm and metaphor: Evidence from eye movements. Journal of Experimental Psychology: Learning, Memory, and Cognition, 42(3), 433.

Pogue, A., Kurumada, C., & Tanenhaus, M. K. (2016). Talker-specific generalization of pragmatic inferences based on under-and over-informative prenominal adjective use. *Frontiers in psychology*, *6*, 2035.

Politzer-Ahles, S., & Fiorentino, R. (2013). The realization of scalar inferences: Context sensitivity without processing cost. *PloS One*, *8*(5), e63943.

R Core Team. (2020). *R: A language and environment for statistical computing* [Manual]. Vienna, Austria: R Foundation for Statistical Computing.

Rees, A., & Bott, L. (2018). The role of alternative salience in the derivation of scalar implicatures. *Cognition*, *176*, 1–14.

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., & Ripley, M. B. (2013). Package "mass." *Cran r*, *538*, 113–120.

Ronai, E., & Xiang, M. (2023). Tracking the activation of scalar alternatives with semantic priming. *Experiments in Linguistic Meaning*, *2*, 229-240.

Ronderos, C. R., Tomlinson, J., & Noveck, I. (2023). *Intentionality, speaker's attitude and the processing of verbal irony*. Presented at the Proceedings of the 42nd Annual Meeting of the Cognitive Science Society.

Scott-Phillips, T. C. (2008). Defining biological communication. *Journal of Evolutionary Biology*, *21*(2), 387–395.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, *32*(1), 3–23.

Sedivy, J. C., Tanenhaus, M., Chambers, C., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147.

Singmann H, Bolker B, Westfall J, Aust F, Ben-Shachar MS, Højsgaard S, Fox J, Lawrence MA, Mertens U, Love J, et al. (2020) Package "afex." https://cran.r-project.org/web/packages/afex/afex.pdf

Spaulding, S. (2020). What is mindreading? *Wiley Interdisciplinary Reviews: Cognitive Science*, *11*(3), e1523.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Spotorno, N., & Noveck, I. A. (2014). When is irony effortful? *Journal of Experimental Psychology: General*, *143*(4), 1649–1665. https://doi.org/10.1037/a0036630

Tomlinson Jr, John M., Bailey, T. M., & Bott, L. (2013). Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of Memory and Language*, *69*(1), 18–35.

Tomlinson Jr, John Michael, & Ronderos, C. R. (2021). Does intonation automatically strengthen scalar implicatures? *Semantics and Pragmatics*, *14*, 4.

van Tiel, B., Noveck, I., & Kissine, M. (2018). Reasoning with "Some." *Journal of Semantics*. https://doi.org/10.1093/jos/ffy012

van Tiel, B., Pankratz, E., & Sun, C. (2019). Scales and scalarity: Processing scalar inferences. *Journal of Memory and Language*, *105*, 93–107. https://doi.org/10.1016/j.jml.2018.12.002

Van Tiel, B., & Schaeken, W. (2017). Processing conversational implicatures: Alternatives and counterfactual reasoning. *Cognitive Science*, *41*, 1119–1154.

Wickham, H. (2016). Ggplot2: Elegant graphics for data analysis. Springer-Verlag New York.

Wickham, H., François, R., Henry, L., & Müller, K. (2020). Dplyr: A grammar of data manipulation [Manual].

Zehr, J., & Schwarz, F. (2018). *PennController for Internet Based Experiments (IBEX)*. https://doi.org/10.17605/OSF.IO/MD832