Sources of cognitive cost in scalar implicature processing: A review

Ahmed Khorsheed[1]*, Jessica Price[2] and Bob van Tiel[3]

[1]Faculty of Modern Languages and Communication, Universiti Putra Malaysia, Serdang, Malaysia, [2]School of Psychology, University of Nottingham Malaysia, Semenyih, Malaysia, [3]Faculty of Philosophy, Theology and Religious Studies, Radboud University, Nijmegen, Netherlands

Research in Experimental Pragmatics has shown that deriving scalar implicatures involves effort and processing costs. This finding was robust and replicated across a wide variety of testing techniques, logical terms, populations, and languages. However, a question that remains disputed in the literature is whether this observed processing cost is a product of the inferential process itself or other logical properties whose computation taxes cognitive resources independently of the inferential mechanism. This paper has two objectives: one is to review the previous experimental work on scalar implicatures and how it evolved in the literature, and the other is to discuss possible factors that render computing scalar implicatures cognitively effortful. Implications and directions for future research are provided.

Introduction

When people speak, their utterances very often do not fully encode what they mean, and the context usually leaves room for a variety of interpretations. One may ask how listeners understand the speaker's intended meaning, how they settle on an interpretation, or how inferential comprehension can ever be achieved. These questions have been the center of several discussions in the literature, especially after the philosopher Paul Grice, in his paper *Logic and Conversation*, introduced the notion of implicatures and distinguished between what is said and what is conversationally implicated (Grice, 1975). Grice's proposal and analysis of the various species of conversational implicature were seminal to the field and they changed the way in which pragmatics was conceived at the time. His work (by providing a framework and vocabulary) led to the start of a plethora of experimental work on scalar implicatures that later enriched our understanding of the cognitive processes and representations involved in utterances interpretation (e.g., Noveck, 2001; Noveck and Posada, 2003; Bott and Noveck, 2004; Breheny et al., 2006, to mention a few). Several of these studies focused on how listeners treat utterances with weak logical terms (e.g., <some, all>, <or, and>, <possible, necessary>), and how the implicature associated with them manifests itself in real-time. To illustrate with one exemplary case how an implicature actually works (i.e., emerges and is entertained), let's imagine that there is a dinner party and Henry tells Jane the following:

1(a) Some of the guests have arrived

(b) All of the guests have arrived

(c) Not all of the guests have arrived

Upon Jane's hearing of (1a), she might well-inquire about what Henry was implying. It is likely that Jane would draw out proposition (1c) [i.e., the negation of (1b)]. Proposition (1c) is not encoded in the meaning of the words that the speaker has uttered, but rather it was worked out by the listener on the basis of what was linguistically encoded in (1a). The speaker's use of *some* in "Some of the Xs" compelled the listener to seek out *not all*, i.e., "Not all of the Xs." Logically speaking, the term *some* is compatible with the logical term *all* (i.e., *some* can be glossed as *some and perhaps all*), but if the speaker had really meant to express *all*, he would have said it, since (1b) is a more informative proposition and would make a greater contribution to the goal of the conversation, and the speaker could have obeyed Grice's first Maxim of Quantity *Make your contribution as informative as is required*. But since he did not, the listener may infer that the speaker believed that proposition (1b) is not true in the first place, and that he chose not to utter (1b) to obey Grice's Supermaxim of Quality *Try to make your contribution one that is true*. According to Grice, the inference from (1a) to (1c) is called a conversational implicature, or a *scalar implicature* as was referred to by an account by Horn (1972) who suggests that scalar implicature derivation draws on pre-existing scales consisting of expressions that vary in informational strength. The speaker's use of a weaker term (e.g., *some*) on a given scale could be taken to implicate that the proposition that would have been expressed by a stronger term (e.g., *all*) on the same scale is not the case.

This Gricean account of implicature derivation has implications for language processing. In other words, the inference from the utterance containing *some* in (1a) to the scalar implicature *not all* in (1c) goes through an evaluation of not only what the speaker said and the context[1], but also of what he might have said but did not. This type of effort-demanding inference makes the Gricean account implausible from a cognitive point of view, that is, Grice did not provide subtle specifics about how scalar implicatures are computed on-line, and thus never meant to advance a processing theory (Noveck and Sperber, 2007; Geurts and Rubio-Fernández, 2015). Therefore, post-Gricean pragmatic theorists who aimed to formulate processing models that are squarely set within the computational view of the mind were mainly split into two groups, those who argue for the idea that scalar implicature derivation is an automatic process that occurs at no cost (i.e., the default account, e.g., Levinson, 2000), and those who argue for more context-dependent and effortful derivation of scalar

___

1 Even single sentence utterances can create their own context through a variety of presupposition triggers and information-structure triggers (Breheny et al., 2006, p. 445).

implicatures (i.e., Relevance theory, e.g., Sperber and Wilson, 1986; Wilson and Sperber, 1998).

Levinson's (2000) default account builds on Grice's edifice of inferential comprehension and treats scalar implicatures as *generalized* conversational implicatures because they, in the absence of special circumstances, are carried by the use of a certain form of words, and therefore are codifiable to some degree. This account is much less concerned with the speaker's intentions (Levinson, 2000, p. 13). Scalar implicatures go through by *default* (i.e., cost-free), and they are only canceled if there are contextual demands to do so. It is only in the cancellation stage, when the default pragmatic meaning needs to be overridden, that time delays and processing costs are likely to occur. The fact that these scalar implicature are generalized by default would add to the speed and efficiency of communication.

Proponents of Relevance theory (e.g., Sperber and Wilson, 1986; Wilson and Sperber, 1998, 2012; Carston, 2002), on the other hand, reject the notion of automaticity in generating scalar implicatures and they instead suggest that making all pragmatic interpretations including scalar implicature depends on a more general principle of relevance and its role in cognition. A linguistic input (i.e., utterance) is considered *relevant* to an individual (or their internal cognitive processes) when it *connects* with available contextual assumptions, or when the inferential process coincides with some consequences that might make the utterance relevant as expected (for discussion, see Noveck and Sperber, 2007). In other words, utterances are pieces of evidence about the speaker's meaning, and appreciating the speaker's meaning is achieved by taking into account the linguistically encoded meaning and its relevance to context. From a relevance-theoretic pragmatic perspective, human cognition is geared to the maximization of relevance in communication; and therefore, the processing of an utterance and the mental effort associated with it are influenced by how deeply the listener is willing to bridge the gap between the linguistic meaning of an utterance and the speaker's meaning (by doing "enrichments, revisions and reorganizations of existing beliefs and plans"). That said, this account views scalar implicature making as a fully-fledged inferential process that occurs with a processing cost, unless the context makes the scalar implicature highly accessible.

The efforts to find experimental data that unravel the implicature process and bear on the analyses of pragmatic theories fundamentally grew out of developmental work on children's *understanding* of scalar implicatures (Noveck, 2001). In a reasoning scenario, Noveck investigated how children and adults would evaluate utterances with weak logical terms (e.g., *might*) when the context indicates that the stronger alternative (e.g., *must*) is the case. In his experiment, using the hidden parrot-and-bear paradigm (see also Noveck et al., 1996), participants saw two boxes, one that contains a parrot (parrot only) and another a parrot and a bear (parrot + bear).

Subsequent to this, a third box that remains covered is shown to participants and they are informed that the contents of this covered box resembles one of the two exposed boxes (i.e., there is necessarily a parrot and possibly also a bear). At this point, the experiment was set up so that a puppet would utter competing modal statements that provide participants with background evidence that generates different truth values (i.e., necessary, possible, non-necessary, and impossible). Participants need to evaluate these statements and determine whether or not they agree with the puppet's statement (e.g., "There has to be a parrot in the box," "There cannot be a bear in the box," etc.). The statement "There might be a parrot in the box" is the critical trial of the experiment because it has two possible interpretations: logical (i.e., it expresses possibility and it is compatible with the deontic meaning of necessity expressed by *must/has to*), and pragmatic (i.e., it is restrictive and not compatible with *must/has to*). Noveck (2001) found that younger children were overwhelmingly logical in responding to underinformative modal statements relative to adults, and that children's tendency to adopt a logical interpretation drops with age. This effect, the pragmatic-enrichment-with-age effect, was consistent and replicated across studies using utterances with other logical terms like *some* (see Experiment 3, Noveck, 2001) *and* (Noveck and Chevaux, 2002), as well as different testing materials (Papafragou and Musolino, 2003; Guasti et al., 2005; Pouscoulous et al., 2007) and different languages (e.g., Katsos et al., 2016). Noveck explained this developmental effect in light of Relevance theory, especially by arguing that scalar implicature generation is not automatic as the default account would suggest, and that children's understanding of scalar implicatures establishes itself on the edifice of reasoning and psychology, rather than grammar or default rules.

Noveck's arguments about implicature understanding among children caught the collective attention of cognitive scientists and set in motion multiple follow-up studies that investigated scalar implicature processing among adults. The question was specifically about whether pragmatic processing would emerge the same way as it does developmentally among children, *viz*, from a semantic (linguistically encoded meaning) to a pragmatically enriched one, or whether the time taken to arrive at a pragmatically enriched meaning is associated with an increased mental effort (i.e., processing cost). Bott and Noveck (2004) were the first to experimentally investigate these questions and provide evidence indicating that going through a pragmatic interpretation is associated with longer (deeper) processing (see also Breheny et al., 2006; De Neys and Schaeken, 2007; Huang and Snedeker, 2009a). Bott and Noveck, using an online sentence judgment task, had French participants read underinformative sentences with *Some* (e.g., "Some cats are mammals"), among other control items like "All cats are mammals" (true), "All mammals are cats" (false), "Some mammals are cats" (true),

etc. Participants are required to evaluate these quantified statements and judge whether they are true or false by pressing a response button. In the underinformative sentence trial, TRUE responses are taken to mean that *Some* was treated as *some and perhaps all* (i.e., logical meaning), whereas FALSE responses are taken to indicate the *some but not all* (i.e., pragmatic) interpretation. Four experiments were conducted and, similar to previous work by Noveck and Posada (2003), general results showed that participants took a significantly longer time to arrive at a pragmatic interpretation than a logical one. This slowdown in processing the scalar implicature was also significantly larger when compared to the time taken to process control trials. Moreover, when the time available for responding was manipulated so that participants were asked to respond in 900 ms (Short Condition) or 3,000 ms (Long Condition) (see Experiment 4), Bott and Noveck found evidence that participants derived fewer scalar implicatures in the *Short* condition than in the *Long* condition (see also Chevallier et al., 2008, for similar evidence with disjunctive statements). This indicated that pragmatic responding is linked with processing effort.

At the time, these initial data which supported the contextual view of Relevance theory were considered so counter-intuitive that they prompted people to come up with more severe tests of its veracity. One worry was that the single-sentence utterances used by Noveck and colleagues were presented in artificial contexts, and thus risked being not generalizable to other tasks, such as sentence processing. Put differently, the effect would be more convincing if it occurred in a more natural context or if it were carried out under more severe testing conditions, and this in itself motivated researchers to engage with a wider variety of techniques. As far as adult processing is concerned, this led to work on scalars using text comprehension vignettes (e.g., Breheny et al., 2006), sentence processing (with eye-tracking tasks, e.g., Huang and Snedeker, 2009a), or with dual-tasks to determine whether one can find costs when memory is taxed (De Neys and Schaeken, 2007). Breheny et al. (2006), for instance, asked participants to read vignettes piecemeal in a self-paced reading task. The experiment involved presenting a piece of the text first, and then other parts of a text would unfurl if a participant manually tapped on a computer keyboard's button. The trigger-containing phrase always occurred at the end of a given discourse. This procedure, phrase-by-phrase reading, allowed Breheny and colleagues to have reading time profiles for the target segment in their experiment(s), which enabled them to test whether a given prediction is supported or not. The discourses given to participants are similar to those in (2a,b) below. The phrase with the disjunctive *or* is read exclusively in (2a) and inclusively in (2b). While Relevance theory predicts longer processing in the upper-bound context (because *or* calls for a pragmatic enrichment), the default account contrastively

argues for longer reading times in the lower-bound context (because the scalar implicature on the disjunctive *or* is generated by default and then subsequently canceled). Note that the context in (2b) makes the reading without the scalar implicature more plausible.

2(a) Upper-bound context: John was taking a university course/and working at the same time./For the exams/he had to study/from short and comprehensive sources./Depending on the course,/he decided to read/the class notes or the summary./

(b) Lower-bound context: John heard that/the textbook for Geophysics/was very advanced./Nobody understood it properly./He heard that/if he wanted to pass the course/he should read/the class notes or the summary./
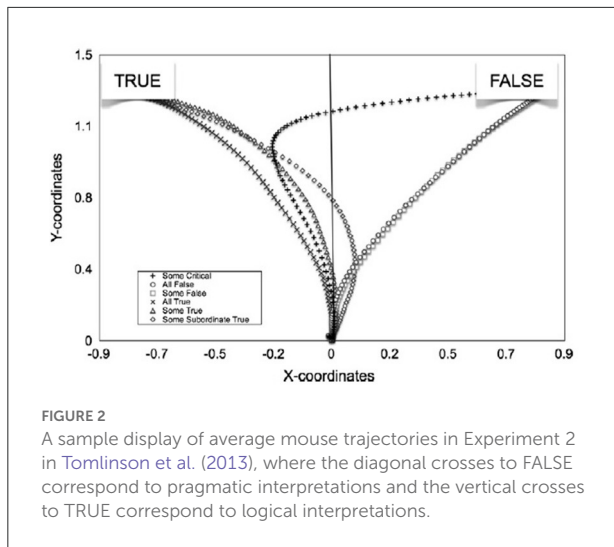
As predicted by Relevance theory, Breheny and colleagues showed that texts that prompted an implicature were associated with a significant processing slowdown (longer reading time) in the upper-bound context relative to the lower-bound context. In other words, when the disjunction *or* occurred in a context where the literal interpretation does not sufficiently satisfy participants' internal thresholds of relevance, more time was needed to derive the scalar implicature *class notes or summary but not both*. However, when the context did not warrant an implicature in the lower-bound context (2b), the disjunction *or* in *the class notes or summary* was compatible with a semantic reading, and thus participants tended to exhibit a faster processing time, which disconfirms the default view (see also Bergen and Grodner, 2012; Breheny et al., 2013).

As far as language comprehension is concerned, experimenters have tested scalar implicatures in visual world eye-tracking paradigms (Huang and Snedeker, 2009a; Grodner et al., 2010), and mouse-tracking paradigms (Tomlinson et al., 2013) on the ground that these techniques provide a more implicit measure of how interpretation unfolds over time prior to overt judgments. Researchers were particularly keen on exploring the real-time interaction between semantic and pragmatic processes during language comprehension, especially by how quickly a listener would isolate objects or a part of the scene as a function of the words in a sentence. For instance, in work by Huang and Snedeker (2009a), participants were prompted to view a podium with four quadrants, in which four pictures of children characters were placed (see Figure 1). The two left quadrants and two right quadrants contain children of the same gender, and each child is paired with a different set of objects. On a sample trial, the two left quadrants contained a boy in each, one with two socks and the other with no objects, whereas the two right quadrants contained a girl in each, one with two socks (pragmatic target) and another with three soccer balls (logical/semantic target). Huang and Snedeker (2009a) acted out a preamble for this

display to create a context against which utterances would be interpreted. This preamble states that a coach gave two socks to one of the boys, two socks to one of the girls, three soccer balls to the other girl, and nothing to the other boy. Then, participants had to follow instructions such as *Point to the girl that has some/all/two/three of the socks/soccer balls.* So, if the speaker says "Point to the girl that has all of the soccer balls", the participant is expected to fixate on the girl who has all of soccer balls. In the target condition, the critical character (the girl) has a subset of one type of item (the socks) while the other has the total set of a second type of item (the soccer balls), and in the target trial participants were asked to *Point to the girl that has some of the soc* (i.e., *SOC*—ks, or *SOC*—cer balls). Huang and Snedeker (2009a) assumed that, if the logical interpretation was to be computed prior to the inference, then the participants, upon hearing the word *some of,* would not be likely to fixate their gaze on the pragmatic target (the girl with two socks) until the referent is mentioned (note that the initial part of the two candidate referents are similar phonologically), since both targets are compatible with the logical interpretation of *some.*

The results obtained from Huang and Snedeker's study were consistent with the proposition that the logical interpretation remains active well after reading "some." For example, they found that the time taken to identify the referent with commands using *all* (e.g., *Point to the girl that has all the soccer balls*) as well as in other control conditions using number (e.g., *Point to the girl that has two/three of the soccer balls*) took ∼200–400 ms. In contrast, for commands with *some*, identification did not occur until 1,000–1,200 ms after the quantifier onset. Huang and Snedeker (2009a) concluded that the pragmatic reading takes additional time to compute relative to the logical meaning, which provides evidence against

FIGURE 2
A sample display of average mouse trajectories in Experiment 2 in Tomlinson et al. (2013), where the diagonal crosses to FALSE correspond to pragmatic interpretations and the vertical crosses to TRUE correspond to logical interpretations.

the default account and the notion of automaticity in scalar implicature making.

In other related experimental work, similar to Huang and Snedeker's, Tomlinson et al. (2013) investigated whether people would first access the basic "logical" meaning before they form the scalar implicature, or whether arriving at the pragmatic meaning is a one-step process that is directly incorporated into the sentence representation. Tomlinson et al. asked participants to evaluate and judge categorical sentences (e.g., *Some elephants are mammals*) using a mouse-tracking technique. In this testing paradigm, participants are asked to initiate the experiment by pressing the start button located at the bottom center of a computer screen. Subsequent to this, a test sentence is presented in the form of words flashed consecutively at the center of the screen. In their paradigm, participants judge the veracity of these test sentences immediately after the onset of the final word of the sentence. For this purpose, participants are asked to click on the TRUE button placed at the top-left of the screen if they consider the sentence true, or to click the FALSE button placed at the top-right if they consider it false (see Figure 2). Typically, the path with which the mouse moves from the bottom of the screen to the target response can tell us about the underlying cognitive processes that drive response formulation and selection. A direct path from the start button toward the target response button is somewhat reflective of an easy one-step process, whereas an indirect path toward the target response button (i.e., deviating toward the competitor response before returning to the target response) is reflective of a more cognitively difficult two-step process. Such a trajectory would suggest that participants initially interpret the sentence literally and only later derive the scalar implicature.

Tomlinson et al. (2013) compared the mouse paths for the pragmatic and logical responses to underinformative

sentences. They found that the mouse path for the pragmatic interpretations of critical sentences deviated substantially toward the competitor response (TRUE) before crossing back over the medial axis toward the target response (FALSE). Fascinatingly, no such deviation was observed for the logical interpretation or literal controls. Tomlinson et al. (2013) suggested that participants' initial mouse movements toward TRUE in the pragmatic condition strongly suggests that a logical meaning is triggered early in processing, whereas a pragmatic interpretation was activated relatively late in processing. These results were consistent with the relevance-theoretic view which assumes that arriving at a pragmatically enriched meaning comes with processing costs.

In another clever attempt to experimentally investigate the automaticity claim of scalar implicatures, De Neys and Schaeken (2007) conducted an experiment in which they had participants evaluate underinformative sentences such as *Some oaks are trees* while their executive cognitive resources were concurrently burdened by a dot memorization task (i.e., memorizing patterns of varying complexity), also called the spatial storage task (Bethell-Fox and Shepard, 1988). De Neys and Schaeken had a prediction: if the scalar implicature is generated by default, and it is the logical interpretation that is effortful, then burdening the cognitive resources should hamper the logical interpretation process; and therefore, participants should provide more pragmatic interpretations under load. However, if the implicature is not automatic, but requires effortful processing, then participants are expected to stick to more logical interpretations under cognitive load. Critically, this latter prediction was confirmed. Results showed that participants gave significantly fewer pragmatic responses when they had to memorize complex patterns than when memorizing easier ones. Moreover, participants' pragmatic response latencies under complex load were significantly longer than the latencies in control load conditions. Interestingly, the latencies for logical interpretations under load effect did not differ from the latencies in literal controls and fillers; and thus, a finding compatible with the idea that making pragmatic interpretations is indeed cognitively demanding.

In fact, with the passing of time, once there was abundant experimental evidence in favor of Relevance Theory, especially from developmental data (e.g., Noveck, 2001; Papafragou and Musolino, 2003; Pouscoulous et al., 2007; Skordos and Papafragou, 2016), adult data (e.g., Bott and Noveck, 2004; Breheny et al., 2006; Dieussaert et al., 2011; Bott et al., 2012; Tomlinson et al., 2013; Heyman and Schaeken, 2015; van Tiel and Schaeken, 2017; Bott and Frisson, 2022), L2 speakers (e.g., Khorsheed et al., 2022; Zhang and Wu, 2022), dual-tasks (e.g., De Neys and Schaeken, 2007; Dieussaert et al., 2011), visual world paradigms (e.g., Huang and Snedeker, 2009a,b), ERPs (e.g., Noveck and Posada, 2003; Barbet and Thierry, 2016; Spychalska et al., 2016), response rates (e.g., Chevallier et al., 2008; Antoniou et al., 2016; Mazzaggio et al., 2021),

and studies using almost a full inventory of lexical scales (Van Tiel et al., 2014) and bare numerals (Spychalska et al., 2019; Noveck et al., 2022), researchers largely abandoned the default view and they became increasingly confident of the idea that scalar implicatures do not arise for free. However, recently, a number of researchers have questioned this conclusion (e.g., Grodner et al., 2010; Marty and Chemla, 2013; Politzer-Ahles and Fiorentino, 2013; Degen and Tanenhaus, 2015; Barbet and Thierry, 2018). These researchers attempted to offer an alternative account for the apparent evidence in favor of what is essentially the Relevance view. This can be seen in, e.g., Grodner et al.'s work. Grodner and colleagues followed up on the Huang and Snedeker (2009a) studies with suggestions stating that the additional time observed in computing the pragmatically enriched meaning could be due to an artifact in the apparatus of their design. They suggested that, if "some of" is phonetically reduced to "summa," and if the Huang and Snedeker (2009a) experiments are set up with instructions that potentially make the contrast between full sets and subsets more prominent, one can minimize the extra processing effects associated with the scalar implicature. Grodner and colleagues made these aforementioned amendments and others in the design (see original work), showing that participants in the "summa" cases appear to be as reactive as "all" cases (but not completely, see their first-half "summa" results, p. 46).

Despite the efforts to experimentally find out factors that may have "conspired" against finding earlier effects of the inference, it remains hard to completely dismiss the relevance-theoretic account. For example, Degen and Tanenhaus (2015) produced a task (gumball machine) that generally prompts "logical" readings (e.g., pragmatic No responses to *some* in the unpartitioned condition were <20%) and one can still find a pattern of pragmatic processing slowdown (see Degen and Tanenhaus, 2015, p. 694). More recently, Barbet and Thierry (2018) attempted to account for the discrepancy in previous context-dependent reading experiments, especially the processing cost observed in reading the *some*-region in Breheny et al. (2006) study (see also Bergen and Grodner, 2012) but not in Politzer-Ahles and Fiorentino (2013) or Hartshorne et al. (2015). They investigated scalar implicatures on single words, including on critical quantified terms, such as "some" using a Stroop task (i.e., the context is maximally neutral, not biased toward an upper-bound or lower-bound reading) and general results did not reveal an N450 effect linked to a pragmatic interpretation, but they found a P600 effect which potentially points to pragmatic processing (for relevant discussion, see Spotorno et al., 2013; Weiland et al., 2014; Spychalska et al., 2016; Noveck, 2018).

All in all, with that being stated in the Introduction, it is clear that the empirical landscape on scalar implicature processing largely supports the relevance-theoretic pragmatic perspective. In other words, scalar implicatures do not arise by default, but rather they are effortful and involve a processing cost, although

the amount of this cost may vary in the degree to which an individual is willing to make the implicature relevant as expected (i.e., bridging the gap between the linguistic utterance meaning and speaker's meaning) (see Noveck and Sperber, 2007 for an alternative explanation), that is, the scalar implicature will be only generated if it meets the listener's internal assumptions of relevance (also see priming experiments, e.g., Skordos and Papafragou, 2016; Rees and Bott, 2018; Bott and Frisson, 2022). In a nutshell, one can say that the developmental data that were considered counter-intuitive at a time have since been replicated and validated across a wider variety of scalar items, testing techniques, populations, and languages, and thus further confirm the view that scalar implicature derivation is effort-demanding. What factor or factors might account for the cognitive cost of scalar implicature derivation is currently an intriguing question that we address in the next section.

Sources of cognitive cost

Thus far, we have reviewed the literature on scalar implicatures, especially how the experimental enterprise of scalar implicatures evolved in the literature and how the processing cost associated with their derivation was robustly validated across multiple experimental scenarios. However, a current question that remains disputed in the literature is the nature of this cognitive cost. For instance, is it all, or a proportion of it, that belongs to the inferential process responsible for generating scalar implicatures? Are there certain linguistic and logical properties that are taxing cognitive resources independently of the inferential mechanism? In this section, we aim to address these questions and discuss how recent experimental work has attempted to explain the underlying source(s) of cognitive cost in processing scalar implicatures.

Semantic complexity

Several authors have suggested that the cognitive effort observed in deriving scalar implicatures is not a product of the inferential process, but rather due to some difficulty inherent in the semantic structure of pragmatic interpretations compared to logical interpretations. For instance, in a sentence like *Some elephants are mammals*, participants may pass through a stage in which they verify if there are elephants that are *not* mammals, as in *some but not all*, compared to a situation in which an overlap between elephants and mammals exists, *as in some and possibly all* (Grodner et al., 2010; Bott et al., 2012; Marty and Chemla, 2013; van Tiel and Schaeken, 2017), and therefore greater informational complexity that may potentially inflate response times independently of implicature derivation. Put differently, this implies that even if the pragmatic interpretation is genuinely generated quickly and by default,

as the default account suggests, this rapid interpretation does not manifest itself in the results due to the time that is allocated to the verification process in assessing the truth-value of the pragmatic interpretation, thus trading off speed for accuracy.

The possibility that the semantic complexity of pragmatic interpretations is what makes their processing take a longer time than the semantically-plain logical interpretations was recently empirically tested by Bott et al. (2012). Bott and colleagues compared the computation of sentences with an implicit pragmatic interpretation, as in *Some* [but not all] *elephants are mammals* (the pragmatic-some condition) against equivalents whose underinformative meaning was made explicit by adding *Only* at the beginning of the given sentence, as in *Only some elephants are mammals* (the only-some condition). Both sentence forms are thought to generate pragmatic interpretations that are equally complex, but only for the sentences in the pragmatic-some condition is this pragmatic interpretation generated as a scalar implicature (Breheny et al., 2006).

Bott et al. (2012) found evidence that there is an inferential cost in deriving the scalar implicature independent of the semantic complexity. More particularly, the accurate pragmatic responding in the *some*-condition begins 140 ms later than the correct logical responding (470 *vs.* 330 ms, respectively, Experiment 1), whereas the accurate pragmatic responding in the *only some*-condition begins 120 ms later than the accurate logical responding (480 *vs.* 360 ms, respectively, Experiment 2), which suggests that the extra time needed to make a response is not due to the verification process involved in processing the semantic complexity of these sentences, but rather due to costs resulting from the inferential processes involved in computing the pragmatic meaning of those underinformative sentences (see also Marty and Chemla, 2013).

However, while the above explanation suggests that the cost demonstrated in reading critical sentences is not overtly a product of the verification process involved in the semantic complexity of underinformative sentences, Tomlinson et al. (2013) argue that the evidence in these studies could be challenged. More specifically, while the bare quantifier "*Some*" may possibly focus one's attention on the referent (i.e., elephants that have trunks), the operator *Only* with the quantifier *some* in "*Only some*" may focus one's attention on the complement (i.e., elephants that do not have trunks) (see also Moxey et al., 2001 for a relevant illustration). As such, the focusing properties associated with the operator "Only" may have made it easier for participants to verify and process the pragmatic meaning associated with "Only some" than the pragmatic meaning associated with the bare quantifier "Some." That said, it remains indecisive and unclear whether the cognitive cost displayed in processing underinformative sentences is due to the verification processes involved in verifying the semantic complexity of upper-informative statements or to the inferential processes involved in analyzing the implicature (see also Discussion in Antoniou et al., 2016).

Disambiguation-related mechanisms

According to Marty and Chemla (2013), it might be true that processing scalar implicatures comes with memory costs, but information about the specific inferential stage at which the cognitive cost is invested is not available. Marty and Chemla proposed that deriving scalar implicatures involves two processing stages: (i) the decision to derive the inference; and (ii) the calculation of stronger alternatives. The question of whether the *decision* to disambiguate the implicature or the calculation of alternatives is what makes the pragmatic task effortful was empirically tested by using ambiguous *vs.* unambiguous quantifiers—*Some vs. Only some*. The difference between both types of quantifiers is this: while the ambiguous *some* in a sentence like *Some elephants are mammals* can give rise to two potential interpretations, i.e., the logical interpretation (some and possibly all) elephants are mammals *vs.* the pragmatic interpretation (some but not all) elephants are mammals, the unambiguous quantifier *only some* in a sentence such as *Only some elephants are mammals* allows only one plain pragmatic meaning, indicating that not all elephants are mammals.

Marty and Chemla (2013) compared participants' performance in the *Some*-condition against the performance in the *Only Some*-condition in a dual-task methodology whose function is to obstruct the appeal to central working memory in conditions in which the burden is high. The prediction was this: if the inferential process is reliant on working memory resources to derive the implicature, then the participants are supposed to exhibit equivalent derivation behavior in the *Some*- and *Only some*-conditions, because they both give rise to the inference *not all,* though in the *Some*-condition this inference is an implicature; in the *Only some*-condition an entailment. However, if the *decision* to compute the implicature is the only subprocess that turns the processing effortful and not the inferential process *per se*, then the participants are expected to have processing difficulties in the *some*-condition but not in the *only some*-condition. Marty and Chemla found evidence that the latter prediction was borne out. Tapping the participants' executive resources was consistent with a significant decrease in deriving the not-all implicatures in the *some*-condition, but no comparable decrease in computing the not-all inferences in the *only some*-condition. Marty and Chemla proposed that the extra memory cost observed in processing the implicature was only a product of the *decision* to derive the implicature rather than to the computation mechanism *per se*.

A similar view that speaks for this conclusion comes from Noveck and Posada (2003) who propose that "false"

responses to underinformative statements are effort-demanding, i.e., they appear to be linked to "decision-related mechanisms." Moreover, Katsos and Bishop (2011) found evidence that children, relative to adults, are less likely to reject infelicitous statement in paradigms that require participants to judge the truth-value of a statement in a binary judgment task, but they are able to make pragmatic inferences in comparable amounts to adults when they are prompted to judge scalar implicatures using a Likert scale task (see also Skordos and Papafragou, 2016; Rees and Bott, 2018). This children-adult behavior discrepancy was mainly attributed to difficulties in children's ability to make/verbalize judgments. Children may default to simple "yes" because binary judgment tasks do not make the relevance of stronger alternatives salient when children assess underinformative sentences, but the 3-point scale task may activate cues of relevance that encourage participants to be more sensitive to the "appropriateness" of the underinformative statement (Skordos and Papafragou, 2016).

Critically, in the same vein, recent work on scalar implicatures among children showed that children's main problem in processing scalar implicatures seems to lie in their understanding of relevance and/or activating *alternatives* (Noveck, 2001; Huang and Snedeker, 2009a; Barner et al., 2011; Degen and Tanenhaus, 2015, 2016; Skordos and Papafragou, 2016; Rees and Bott, 2018), a process that is thought to involve working memory resources (Chierchia et al., 2001; Reinhart, 2004; Pouscoulous et al., 2007; Gotzner and Spalek, 2017). However, Katsos and Bishop (2011) argue that the process of activating and comparing alternatives is not memory taxing, and rather it is the decision that makes the inferential process effortful (see also Marty and Chemla, 2013 above). That said, future studies may need to determine the underlying cause facilitating scalar implicature processing in ternary judgement tasks compared to binary judgement tasks. Interestingly, recent work by Jasbi et al. (2019) showed that adults, similar to children, prefer intermediate response options in ternary judgment tasks, and hence an indirect piece of evidence standing against the proposition that childrens' main problem in processing scalar implicatures lies at the interplay between response options and activating alternatives. Preferring an intermediate response option, which can manifest itself through avoiding a response option at the positive end of the scale (primarily among children) and negative end of the scale (primarily among adults), still seems to be generic among different types of populations and not related to children participants only. Therefore, we recommend that future work place some focus on the relationship between working memory capacity and implicature derivation among children. Children's difficulty to activate alternatives could be an artifact of limited processing resources (see Chierchia et al., 2001; Reinhart, 2004; Pouscoulous et al., 2007).

Embedded negation

The topic of negation and its processing cost has been the center of many linguistic and psycholinguistic discussions in the literature. Early experimental work on negation, including the explicit negative with *not* and the implicit negative in quantifiers such as *few* or *scarcely any* has shown that integrating negation into the sentence meaning is accompanied with high error rates and longer processing times compared to affirmative sentences (Wason, 1959, 1961; Just and Carpenter, 1971; Clark and Chase, 1972; Moxey et al., 2001; Prado and Noveck, 2006; Deschamps et al., 2015). Horn (1989) also, in his large volume of linguistic work on negation processing, suggests that the concept of negation itself is a multifaceted phenomenon that has complex and systematic interaction with logical operators, especially quantitative scales. On comprehending the processing slowdown associated with pragmatic reading of the quantifier *some*, Rips (1975) suggested that negation could be the source of this slowdown as *Some* may function as *Some-but-not-All* (see also Bott and Noveck, 2004 for a similar argument). Arguably, if this claim is indeed true, this implies that the cognitive cost associated with computing the scalar inference is not a reflection of the inferential process responsible for generating the scalar inference *per se*, but rather of the embedded negation in the meaning of *some-but-not-all* for *some*.

Interestingly, in a recent attempt to empirically test if the inferential process itself or the negation embedded in computing the pragmatic meaning of *some* (not-all) is what makes scalar implicature processing computationally complex, van Tiel et al. (2019) used scalar words that differed in their scalarity, i.e., *positive* scalar words (<or, and>, <might, must>, <some, all>, <most, all>, and <try, succeed>) *vs. negative* scalar words (<low, empty>, <scarce, absent>). According to van Tiel and colleagues, the scalarity of a word determines the polarity of the corresponding scalar inference, i.e., the positive scalar word *some* gives rise to the negative scalar inference *not all*, whereas the negative scalar word *not all* gives rise to the positive scalar inference *not-none*, or, equivalently, to *some*. For instance, the positive scalar word *Some* in the sentence "I ate some of the pie" implies the negative scalar inference "not all the pie," whereas the negative scalar word in the sentence "I did not eat all of the pie" implies the positive proposition that I did not eat none of the pie, i.e., I ate some. van Tiel et al. argued that if the process of scalar inference reading is cognitively demanding, then this cognitive demand should generalize to the entire family of scalar words, no matter whether this is a positive scalar word or a negative scalar word. However, if the positive scalar words— but not the negative scalar ones—are the only scalar words that incur processing costs, then this would mean that the processing cost associated with them is a function of negation embedded in the underlying structure of the pragmatic proposition and

not due to the inferential process responsible for generating the scalar inference.

van Tiel et al. examined the processing of both positive and negative scalar words and found that rejecting the underinformative meaning triggered by the positive scalar words "might," "some," "or," and "most" was consistently associated with processing slowdowns, whereas rejecting the underinformative meaning in the negative scalar words "low" and "scarce"—as well as for the positive scalar word "try" in Experiment 1 only—was made without any noticeable cognitive costs, and therefore the processing costs did not generalize to the entire family of scalar words. On the ground of these results, van Tiel and colleagues argued that the source of processing slowdown in deriving scalar implicatures does not seem to stem from the process that computes the implicature itself, but rather from negation embedded in the pragmatic reading of the positive scalar term. This negation taxes working memory, and therefore, artificially inflates the time taken to make pragmatic inferences (also in line with Cremers and Chemla, 2014; Romoli and Schwarz, 2015).

It is worth noting, however, that the negative lexical scales that were used in van Tiel et al.'s study were exclusively adjectival (e.g., low, scarce), whereas the positive lexical scales were a mix of other parts of speech (i.e., <or, and>, <might, must>, <some, all>, <most, all>, and <try, succeed>). That said, the implicatures on adjectives are triggered differently from those on other lexical scales (Baker et al., 2009; Van Tiel et al., 2014; Gotzner et al., 2018; van Tiel and Pankratz, 2021), and hence the study lacked some sort of control in sampling. Second, we propose that both types of scales (i.e., positive scalar words and negative scalar words), as a function of pragmatic enrichment, involve a scale reversal step in which the stonger scalemate is represented and then denied, i.e., some (∼not all), low (∼not-empty, existence of something). That said, the processing slowdown manifested in reading the *positive* scalar words, say *Some*, may not be overly due to the denial[2] of the stronger scalemate (All). However, the directionality/polarity of the inferred proposition might be of cognitive relevance. More specifically, the interplay between the polarity of these inferred propositions (i.e., negative *vs.* positive polarity) and *possibly* other logical properties (discussed below) could be a significant source of cognitive cost in scalar implicature processing, and potentially the underlying cause of discrepancy between processing positive scalar words and negative scalar words, respectively. Further details about our proposal are given in the following section.

_____

2 Denial refers to embedded negation. For example, in a sentence such as *Some cats meow*, the word *Some* serves to deny a stronger proposition with *All* (i.e., All cats meow) after being mentally represented.

Negation, polarity, monotonicity, and processing

The topic of quantification and quantifier interpretation has been of strong interest to many philosophers, linguists, and logicians. There is a prodigious number of studies that have attempted to explain how quantified statements are interpreted and what logical properties render them computationally more complex (Just and Carpenter, 1971; Horn, 1972; Barwise and Cooper, 1981; Sanford et al., 1996; Paterson et al., 1998, 2009; Moxey et al., 2001; Geurts and van der Slik, 2005; Prado and Noveck, 2006; Geurts et al., 2010; Penka and Zeijlstra, 2010; Penka, 2011; Szymanik and Zajenkowski, 2013; Orenes et al., 2016; Haase et al., 2019). The authors of these studies agree that statements quantified by negative quantifiers are harder to process and are more error prone in comprehension compared to statements quantified by positive quantifiers. More recent research about the processing complexity of negative quantifiers showed that both (i) negative polarity and (ii) downward monotonicity are two logical properties that make the processing of negative quantifiers cognitively more complex relative to their positive antonyms (Agmon et al., 2019).

Negative polarity can be thought of as an operator triggering the "less than" computation on a linear scale (i.e., *below* some standard on a scale) (Agmon et al., 2019, 2021) and, when analyzed morphosyntactically, they contain a hidden negation in their underlying structure (Hackl, 2000; Penka, 2011) that evokes an interaction of the polarity effect with the truth-value of the sentence (Just and Carpenter, 1971; Grodzinsky et al., 2018; Agmon et al., 2019). Downward monotonicity, however, can be thought of as a logical/mathematical property whose environment serves to signal downward entailing contexts toward an empty-set scenario (Agmon et al., 2019; Bott et al., 2019). To illustrate, these negative quantifiers and expressions such as *few, a few, little,* and *a small number* equivalently involve a "less than" computation on a linear scale, and possibly a hidden negation, but the difference is this: while the words *few* and *little* are downward monotone items that license NPIs, such as *any, ever,* or *at all,* the hallmark of downward monotonicity (Ladusaw, 1980), the words *a few* and *a small number* are non-monotone items that are introduced by the article "a" whose existential force does not allow licensing downward monotoncity. In a recent attempt to disentangle the processing complexity effects that are due to negative polarity from the processing complexity effects that are due to downward monotonicity, Agmon et al. (2019) investigated the processing difference between NPIs that are downward monotone (i.e., *few, little*) and NPIs that are non-monotone (i.e., *a few, a small number*) and the results revealed that downward monotonicity adds significant cognitive costs to processing negative quantifiers independent of the negative polarity effect.

As shown and discussed above, the cognitive cost observed in computing scalar implicatures triggered by the use of weak scalar quantifiers/words is currently disputed in the literature, especially whether all, or a proportion of it, is an artifact of the semantic complexity of underinformative sentences, disambiguation-related mechanisms (e.g., Noveck and Posada, 2003; Marty and Chemla, 2013), negation embedded in the implicature meaning (e.g., Bott and Noveck, 2004; van Tiel et al., 2019), engagement in Theory of Mind (ToM) reasoning (e.g., Apperly et al., 2006; Van Tiel and Kissine, 2018; Fairchild and Papafragou, 2021), executing the epistemic state (e.g., Bergen and Grodner, 2012; Fairchild and Papafragou, 2021), and/or contrasting and evaluating alternatives (e.g., Degen and Tanenhaus, 2016; Skordos and Papafragou, 2016; Gotzner and Romoli, 2018; Rees and Bott, 2018). Interestingly, recent work by van Tiel et al. (2019) suggests that a large majority of the previous work on scalar implicatures have examined scalar implicature processing using almost exclusively positive scalar words that give rise to negative propositions, and hence the cognitive cost observed in computing scalar implicatures (at least for studies with *some*) could be an artifact of the negation embedded in mentalizing the implicature *Not-all*. More specifically, van Tiel et al. proposed that negative quantifiers give rise to positive pragmatic propositions that are easy to process, whereas positive quantifiers give rise to pragmatic propositions embedded with negation whose computation renders scalar implicature comprehension more complex. For instance, in a sentence such as *some elephants are mammals*, the use of the positive quantifier *some* can pragmatically entail that *not all elephants are mammals*. The denial of the stronger scalemate *all* for the sake of enriching *some* is what contributes to a heavier cognitive load. Conversely, negative quantifiers (e.g., few) usually stimulate positive propositions (i.e., not-none, equivalent to *some*), and this positivity in the proposition is thought to make them computationally less complex compared to their negative counterparts generated by positive quantifiers, and therefore rapid processing times.

In this review paper, we suggest that negation could be a core cognitive step in pragmatic computation, but we propose that negation may not be the exclusive property responsible for the processing slowdown of scalar implicatures triggered by positive scalar words (e.g., some). It is possible that both types of lexical scales (positive scalar words and negative scalar words) may involve, as a function of pragmatic enrichment, a scale reversal step in which the stonger scalemate that holds with the truth of the quantified statement was mentally represented and then denied, i.e., *some-but-not-All, few-but-not-None,* respectively (see also the presupposition-denial account for Moxey et al., 2001; Moxey, 2006), and therefore it is not clear if embedded negation is the source of slowdown in processing scalar implicatures triggered by positive scalar words (but not negative scalar words). However, we suggest that the directionality/polarity of the inferred proposition is possibly of cognitive relevance: the positive inferred proposition points away from zero toward the ordinary way we perceive the world (Clark, 1973), whereas the negative inferred proposition points in a less natural direction, toward the zero end, that makes it harder to process (De Soto et al., 1965; McGonigle and Chalmers, 1996; Geurts and van der Slik, 2005; Hoorens and Bruckmüller, 2015; Agmon et al., 2019).

According to Geurts and van der Slik (2005), we say that expression X is upward or downward entailing to denote the inference licensed by X, but we say that an inference moves upward or downward when we are describing the *actual* inference. This suggestion is in line with the tenets of the scalarity-based account by van Tiel et al. (2019) who also suggest that pragmatic inferencing moves in the opposite direction of the semantics of the quantifier, and hence positive quantifiers and negative quantifiers are likely to trigger different logical effects whose computation would result in different processing costs. These accounts which suggest that the entailment of the quantifiers can move upward or downward depending on the context seem to dovetail with the Focus account by Sanford et al. (1996) who studied the different focus effects of positive and negative quantifiers and proposed that (a) quantifier interpretation relies on attributions of a speaker's expectations and that (b) quantifier interpretation is context-dependent (see also Ingram et al., 2014). Their account proposes that the entailment of a quantifier can move upward or downward depending on whether the inference is calculated or not. This was also faithfully explained by Bott and Noveck who investigated scalar implicatures triggered by the positive quantifier *some* in sentences such as *Some elephants are mammals*. They proposed that participants who choose to interpret *Some* logically are likely to put their focus on the reference set, whereas those who choose to look for a more relevant reading may place their focus on the Complement set, i.e., whether there is a set of non-mammalian elephants, and therefore reading *some* as *some-but-not-all via* a pragmatic operation that transforms the positive *some* into a more complex negative quantifier whose interpretation becomes monotone-decreasing that licenses a compset scenario in which these non-mammalian elephants, after searching, are not existent (i.e., null-Complement set). This null-complement set effect is possibly in tandem with the empty-set effect triggered by monotone-decreasing quantifiers in Bott et al. (2019).

Bott et al. (2019) examined the empty-set effect in an experiment that required the participants to evaluate the truth of a negatively quantified statement such as *less than five dots are blue* in the face of a 0-model picture that does not have the target color dots. They found that (i) rejecting these downward-entailing inferences is more error prone in comprehension, and that (ii) the specification of an algorithm for evaluating empty-set situations was consistent with processing delays

compared to quantifiers that do not trigger empty-set situations. Critically, Bott and colleagues also showed evidence that the cognitive effort involved in processing empty-set evaluations was independent of the cognitive effort needed to compute the scalar implicature *fewer-than-five-but-not-None.* That said, this evidence indirectly suggests that the intrinsic nature of the cognitive cost observed in processing the implicature *Not-all* in previous studies may not be overly related to the inferential process, but rather to other processes that may involve evaluating empty-set scenarios *embedded* in a downward-entailing environment (see also Chemla et al., 2011 for their discussion of perceived *vs.* formal monotonicity). Nevertheless, this proposition remains speculative and can only be confirmed if it is experimentally tested in future work.

All in all, and as the evidence shows in these experimental illustrations, negative polarity, downward monotonicity, and empty-set evaluation are three logical variables whose computation seems to contribute independent proportions of cost in processing negative quantifiers/inferences in pragmatics. According to Agmon et al. (2019), monotonicity contributes about 30% of negative polarity, but Bott et al. (2019) suggest that it is the empty-set effect and not downward monotonicity that contributes to the processing cost of monotone-decreasing quantifiers. Taken together, these variables seem to suggest that downward-entailing inferences are computationally more complex than their upward-entailing counterparts, but the question of how negative pragmatic inferences are computed remains speculative amidst these confounding variables. While our discussion of these variables was extremely sketchy, we hope that the nature of our argument is clear. Future research may need to disentangle these seemingly interrelated proportions of cognitive cost involved in downward-entailing inferences. Is the cognitive cost involved in computing negative inferences triggered by positive scalar words solely related to the inferential process, or it is all, or a proportion of it, a reflection of the cost involved in processing other computational properties, such as negative polarity, downward monotonicity, and empty-set scenario evaluation? It may also pay if future work focus on negative scalar words that give rise to positive upward entailing propositions, and thus the variables negative polarity, downward monotonicity, and empty-set effects can be controlled for[3]. Most previous studies focused on *positive* scalar words that give rise to *negative* propositions; and therefore, there is still a scarcity of work pinning the processing difference between scalar implicatures triggered by positive and negativ*e scalar words.

Conclusion

This review has discussed the experimental record on scalar implicatures and the processing cost associated with their derivation. As discussed in the Introduction, the presence of this processing cost is robust, replicable, and has enjoyed large empirical support from multiple experimental scenarios, including data obtained from children, adults, experiments using visual world eye-tracking paradigms, mouse-tracking paradigms, ERPs, judgment tasks, and reading comprehension vignettes, as well as from different testing material (logical terms, adjectives, bare numerals) and languages. However, some researchers have questioned this conclusion, especially whether or not this observed processing cost corresponds to the inferential process itself or other computational properties that may artificially inflate processing. Our review has put the spotlight on these variables that are thought to contribute to processing scalar implicatures, which include the semantic complexity of some test material, the decision to disambiguate the implicature, the embedded negation in the inferential process, and other semantic and logical properties that relate to negative polarity, downward monotonicity, and empty-set scenarios whose role in the cost observed in scalar implicature processing is still unknown. With that being said, we recommend that future work focus on testing scalar implicatures while taking into account these said caveats so as to provide useful insights into the factors that contribute to scalar implicature processing. In so doing, one can gain important insights into the theories and phenomena related to scalar implicature. This would benefit Experimental Pragmatics specifically as well as scalar implicature research more generally.

Author contributions

AK wrote the first draft of the manuscript. All authors contributed to the intellectual property of the manuscript and they read, reviewed, and approved the submitted version.

Acknowledgments

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

---

3   It is worth noting that some monotone increasing quantifiers can also license empty-set situations (Bott et al., 2019, pp. 11–12).

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Agmon, G., Bain, J. S., and Deschamps, I. (2021). Negative polarity in quantifiers evokes greater activation in language-related regions compared to negative polarity in adjectives. *Exp. Brain Res.* 239, 1427–1438. doi: 10.1007/s00221-021-06067-y

Agmon, G., Loewenstein, Y., and Grodzinsky, Y. (2019). Measuring the cognitive cost of downward monotonicity by controlling for negative polarity. *Glossa J. Gen. Linguist.* 4, 1–18. doi: 10.5334/gjgl.770

Antoniou, K., Cummins, C., and Katsos, N. (2016). Why only some adults reject under-informative utterances. *J. Pragmat.* 99, 78–95. doi: 10.1016/j.pragma.2016.05.001

Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C., and Samson, D. (2006). Is belief reasoning automatic? *Psychol. Sci.* 17, 841–844. doi: 10.1111/j.1467-9280.2006.01791.x

Baker, R., Doran, R., McNabb, Y., Larson, M., and Ward, G. (2009). On the non-unified nature of scalar implicature: an empirical investigation. *Int. Rev. Pragmat.* 1, 211–248. doi: 10.1163/187730909X12538045489854

Barbet, C., and Thierry, G. (2016). Some alternatives? Event-related potential investigation of literal and pragmatic interpretations of some presented in isolation. *Front. Psychol.* 7:1479. doi: 10.3389/fpsyg.2016.01479

Barbet, C., and Thierry, G. (2018). When some triggers a scalar inference out of the blue. An electrophysiological study of a Stroop-like conflict elicited by single words. *Cognition* 177, 58–68. doi: 10.1016/j.cognition.2018.03.013

Barner, D., Brooks, N., and Bale, A. (2011). Accessing the unsaid: the role of scalar alternatives in children's pragmatic inference. *Cognition* 118, 84–93. doi: 10.1016/j.cognition.2010.10.010

Barwise, J., and Cooper, R. (1981). Generalized quantifiers and natural language. *Linguist. Philos.* 4, 159–219. doi: 10.1007/BF00350139

Bergen, L., and Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *J. Exp. Psychol. Learn. Mem. Cogn.* 38:1450. doi: 10.1037/a0027850

Bethell-Fox, C. E., and Shepard, R. N. (1988). Mental rotation: effects of stimulus complexity and familiarity. *J. Exp. Psychol. Hum. Percept. Perform.* 14, 12–23. doi: 10.1037/0096-1523.14.1.12

Bott, L., Bailey, T. M., and Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *J. Mem. Lang.* 66, 123–142. doi: 10.1016/j.jml.2011.09.005

Bott, L., and Frisson, S. (2022). Salient alternatives facilitate implicatures. *PLoS ONE* 17:e0265781. doi: 10.1371/journal.pone.0265781

Bott, L., and Noveck, I. A. (2004). Some utterances are underinformative: the onset and time course of scalar inferences. *J. Mem. Lang.* 51, 437–457. doi: 10.1016/j.jml.2004.05.006

Bott, O., Schlotterbeck, F., and Klein, U. (2019). Empty-set effects in quantifier interpretation. *J. Semant.* 36, 99–163. doi: 10.1093/jos/ffy015

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

Breheny, R., Katsos, N., and 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, 434–463. doi: 10.1016/j.cognition.2005.07.003

Carston, R. (2002). *Thoughts and Utterances: The Pragmatics of Explicit Communication.* Oxford: Blackwell. doi: 10.1002/9780470754603

Chemla, E., Homer, V., and Rothschild, D. (2011). Modularity and intuitions in formal semantics: the case of polarity items. *Linguist. Philos.* 34, 537–570. doi: 10.1007/s10988-012-9106-0

Chevallier, C., Noveck, I. A., Nazir, T., Bott, L., Lanzetti, V., and Sperber, D. (2008). Making disjunctions exclusive. *Q. J. Exp. Psychol.* 61, 1741–1760. doi: 10.1080/17470210701712960

Chierchia, G., Crain, S., Guasti, M. T., Gualmini, A., and Meroni, L. (2001). "The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures," in *Proceedings of the 25th Boston University Conference on Language Development*, Vol. 25 (Boston, MA), 157–168.

Clark, H. H. (1973). "Space, time, semantics, and the child," in *Cognitive Development and Acquisition of Language,* (New York, NY: Academic Press) 27–63. doi: 10.1016/B978-0-12-505850-6.50008-6

Clark, H. H., and Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cogn. Psychol.* 3, 472–517. doi: 10.1016/0010-0285(72)90019-9

Cremers, A., and Chemla, E. (2014). "Direct and indirect scalar implicatures share the same processing signature," in *Pragmatics, Semantics and the Case of Scalar Implicatures,* ed. S. P. Reda (London: Palgrave Macmillan UK) 201–227. doi: 10.1057/9781137333285_8

De Neys, W., and Schaeken, W. (2007). When people are more logical under cognitive load dual task impact on scalar implicature. *Exp. Psychol.* 54, 128–133. doi: 10.1027/1618-3169.54.2.128

De Soto, C. B., London, M., and Handel, S. (1965). Social reasoning and spatial paralogic. *J. Pers. Soc. Psychol.* 2, 513–513. doi: 10.1037/h0022492

Degen, J., and Tanenhaus, M. K. (2015). Processing scalar implicature A constraint-based approach. *Cogn. Sci.* 39, 667–710. doi: 10.1111/cogs.12171

Degen, J., and Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: a visual world eye-tracking study. *Cogn. Sci.* 40, 172–201. doi: 10.1111/cogs.12227

Deschamps, I., Agmon, G., Loewenstein, Y., and Grodzinsky, Y. (2015). The processing of polar quantifiers, and numerosity perception. *Cognition* 143, 115–128. doi: 10.1016/j.cognition.2015.06.006

Dieussaert, K., Verkerk, S., Gillard, E., and Schaeken, W. (2011). Some effort for some: further evidence that scalar implicatures are effortful. *Q. J. Exp. Psychol.* 64, 2352–2367. doi: 10.1080/17470218.2011.588799

Fairchild, S., and Papafragou, A. (2021). The role of executive function and theory of mind in pragmatic computations. *Cogn. Sci.* 45:e12938. doi: 10.1111/cogs.12938

Geurts, B., Katsos, N., Cummins, C., Moons, J., and Noordman, L. (2010). Scalar quantifiers: logic, acquisition, and processing. *Lang. Cogn. Process.* 25, 130–148. doi: 10.1080/01690960902955010

Geurts, B., and Rubio-Fernández, P. (2015). Pragmatics and processing. *Ratio* 28, 446–469. doi: 10.1111/rati.12113

Geurts, B., and van der Slik, F. (2005). Monotonicity and processing load. *J. Semant.* 22, 97–117. doi: 10.1093/jos/ffh018

Gotzner, N., and Romoli, J. (2018). The scalar inferences of strong scalar terms under negative quantifiers and constraints on the theory of alternatives. *J. Semant.* 35, 95–126. doi: 10.1093/jos/ffx016

Gotzner, N., Solt, S., and Benz, A. (2018). Scalar diversity, negative strengthening, and adjectival semantics. *Front. Psychol.* 9, 191–203. doi: 10.3389/fpsyg.2018.01659

Gotzner, N., and Spalek, K. (2017). "The connection between focus and implicatures: investigating alternative activation under working memory load," in *Linguistic and Psycholinguistic Approaches on Implicatures and Presuppositions,* eds S. Pistoia-Reda and F. Domaneschi (Cham: Springer International Publishing), 175–198. doi: 10.1007/978-3-319-50696-8_7

Grice, H. P. (1975). "Logic and conversation," in *Syntax and Semantics 3: Speech Acts,* eds P. Cole and J. L. Morgan (New York, NY: Academic Press), 41–58. doi: 10.1163/9789004368811_003

Grodner, D. J., Klein, N. M., Carbary, K. M., and Tanenhaus, M. K. (2010). "Some," and possibly all, scalar inferences are not delayed: evidence for immediate pragmatic enrichment. *Cognition* 116, 42–55. doi: 10.1016/j.cognition.2010.03.014

Grodzinsky, Y., Agmon, G., Snir, K., Deschamps, I., and Loewenstein, Y. (2018). The processing cost of downward entailingness: the representation and

verification of comparative constructions. *Proc. Sinn Bedeutung 22* 1, 435–451. doi: 10.21248/zaspil.60.2018.475

Guasti, T. M., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., and Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Lang. Cogn. Process.* 20, 667–696. doi: 10.1080/01690960444000250

Haase, V., Spychalska, M., and Werning, M. (2019). Investigating the comprehension of negated sentences employing world knowledge: an event-related potential study. *Front. Psychol.* 10:2184. doi: 10.3389/fpsyg.2019.02184

Hackl, M. (2000). *Comparative Quantifiers.* Cambridge, MA: Massachusetts Institute of Technology.

Hartshorne, J. K., Snedeker, J., Liem Azar, S. Y.-M., and Kim, A. E. (2015). The neural computation of scalar implicature. *Lang. Cogn. Neurosci.* 30, 620–634. doi: 10.1080/23273798.2014.981195

Heyman, T., and Schaeken, W. (2015). Some diferences in some: examining variability in the interpretation of scalars using latent class analysis. *Psychol. Belg.* 55, 1–18. doi: 10.5334/pb.bc

Hoorens, V., and Bruckmüller, S. (2015). Less is more? Think again! A cognitive fluency-based more-less asymmetry in comparative communication. *J. Pers. Soc. Psychol.* 109, 753–766. doi: 10.1037/pspa0000032

Horn, L. R. (1972). *On the Semantic Properties of Logical Operators in English.* Los Angeles, CA: University of California.

Horn, L. R. (1989). *A Natural History of Negation.* Chicago, IL: University of Chicago Press.

Huang, Y. T., and Snedeker, J. (2009a). Online interpretation of scalar quantifiers: insight into the semantics-pragmatics interface. *Cogn. Psychol.* 58, 376–415. doi: 10.1016/j.cogpsych.2008.09.001

Huang, Y. T., and Snedeker, J. (2009b). Semantic meaning and pragmatic interpretation in 5-year-olds: evidence from real-time spoken language comprehension. *Dev. Psychol.* 45:1723. doi: 10.1037/a0016704

Ingram, J., Hand, C. J., and Moxey, L. M. (2014). Processing inferences drawn from the logically equivalent frames half full and half empty. *J. Cogn. Psychol.* 26, 799–817. doi: 10.1080/20445911.2014.956747

Jasbi, M., Waldon, B., and Degen, J. (2019). Linking hypothesis and number of response options modulate inferred scalar implicature rate. *Front. Psychol.* 10, 1–14. doi: 10.3389/fpsyg.2019.00189

Just, M. A., and Carpenter, P. A. (1971). Comprehension of negation with quantification. *J. Verbal Learn. Verbal Behav.* 10, 244–253. doi: 10.1016/S0022-5371(71)80051-8

Katsos, N., and Bishop, D. V. M. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition* 120, 67–81. doi: 10.1016/j.cognition.2011.02.015

Katsos, N., Cummins, C., Ezeizabarrena, M. J., Gavarró, A., Kraljević, J. K., Hrzica, G., et al. (2016). Cross-linguistic patterns in the acquisition of quantifiers. *Proc. Natl. Acad. Sci. U. S. A.* 113, 9244–9249. doi: 10.1073/pnas.1601341113

Khorsheed, A., Md. Rashid, S., Nimehchisalem, V., Geok Imm, L., Price, J., and Ronderos, C. R. (2022). What second-language speakers can tell us about pragmatic processing. *PLoS ONE* 17:e0263724. doi: 10.1371/journal.pone.0263724

Ladusaw, W. A. (1980). On the notion affective in the analysis of negative-polarity items. *J. Linguist. Res.* 1, 1–16.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature.* Cambridge, MA: The MIT Press. doi: 10.7551/mitpress/5526.001.0001

Marty, P. P., and Chemla, E. (2013). Scalar implicatures: working memory and a comparison with only. *Front. Psychol.* 4, 1–12. doi: 10.3389/fpsyg.2013.00403

Mazzaggio, G., Panizza, D., and Surian, L. (2021). On the interpretation of scalar implicatures in first and second language. *J. Pragmat.* 171, 62–75. doi: 10.1016/j.pragma.2020.10.005

McGonigle, B., and Chalmers, M. (1996). "The ontology of order," in *Critical Readings on Piaget,* ed L. Smith (London: Routledge), 279–311.

Moxey, L. M. (2006). Effects of what is expected on the focussing properties of quantifiers: a test of the presupposition-denial account. *J. Mem. Lang.* 55, 422–439. doi: 10.1016/j.jml.2006.05.006

Moxey, L. M., Sanford, A. J., and Dawydiak, E. J. (2001). Denials as controllers of negative quantifier focus. *J. Mem. Lang.* 44, 427–442. doi: 10.1006/jmla.2000.2736

Noveck, I. (2018). *Experimental Pragmatics: The Making of a Cognitive Science.* Cambridge: Cambridge University Press. doi: 10.1017/9781316027073

Noveck, I., and Chevaux, F. (2002). "The pragmatic development of and," in *Proceedings of the 26th Annual Boston University Conference on Language Development,* eds B. Skarabella (Somerville, MA: Cascadilla Press).

Noveck, I., Fogel, M., Van Voorhees, K., and Turco, G. (2022). When eleven does not equal 11: investigating exactness at a number's upper bound. *PLoS ONE* 17:e0266920. doi: 10.1371/journal.pone.0266920

Noveck, I., and Sperber, D. (2007). "The why and how of experimental pragmatics: the case of 'scalar inferences,'" in *Advances in Pragmatics,* ed N. Burton-Roberts (Basingstoke: Palgrave), 184–212. doi: 10.1057/978-1-349-73908-0_10

Noveck, I. A. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition* 78, 165–188. doi: 10.1016/S0010-0277(00)00114-1

Noveck, I. A., Ho, S., and Sera, M. (1996). Children's understanding of epistemic modals. *J. Child Lang.* 23, 621–643. doi: 10.1017/S0305000900008977

Noveck, I. A., and Posada, A. (2003). Characterizing the time course of an implicature: an evoked potentials study. *Brain Lang.* 85, 203–210. doi: 10.1016/S0093-934X(03)00053-1

Orenes, I., Moxey, L., Scheepers, C., and Santamaría, C. (2016). Negation in context: evidence from the visual world paradigm. *Q. J. Exp. Psychol.* 69, 1082–1092. doi: 10.1080/17470218.2015.1063675

Papafragou, A., and Musolino, J. (2003). Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition* 86, 253–282. doi: 10.1016/S0010-0277(02)00179-8

Paterson, K. B., Filik, R., and Moxey, L. M. (2009). Quantifiers and discourse processing. *Linguist. Lang. Comp.* 3, 1390–1402. doi: 10.1111/j.1749-818X.2009.00166.x

Paterson, K. B., Sanford, A. J., Moxey, L. M., and Dawydiak, E. (1998). Quantifier polarity and referential focus during reading. *J. Mem. Lang.* 39, 290–306. doi: 10.1006/jmla.1998.2561

Penka, D. (2011). *Negative Indefinites.* Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199567263.001.0001

Penka, D., and Zeijlstra, H. (2010). Negation and polarity: an introduction. *Nat. Lang. Linguist. Theory* 28, 771–786. doi: 10.1007/s11049-010-9114-0

Politzer-Ahles, S., and Fiorentino, R. (2013). The realization of scalar inferences: context sensitivity without processing cost. *PLoS ONE* 8:e63943. doi: 10.1371/journal.pone.0063943

Pouscoulous, N., Noveck, I. A., Politzer, G., and Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Lang. Acquis.* 14, 347–375. doi: 10.1080/10489220701600457

Prado, J., and Noveck, I. A. (2006). How reaction time measures elucidate the matching bias and the way negations are processed. *Think. Reason.* 12, 309–328. doi: 10.1080/13546780500371241

Rees, A., and Bott, L. (2018). The role of alternative salience in the derivation of scalar implicatures. *Cognition* 176, 1–14. doi: 10.1016/j.cognition.2018.02.024

Reinhart, T. (2004). The processing cost of reference set computation: acquisition of stress shift and focus. *Lang. Acquis.* 12, 109–155. doi: 10.1207/s15327817la1202_1

Rips, L. J. (1975). Quantification and semantic memory. *Cogn. Psychol.* 7, 307–340. doi: 10.1016/0010-0285(75)90014-6

Romoli, J., and Schwarz, F. (2015). "An experimental comparison between presuppositions and indirect scalar implicatures," in *Experimental Perspectives on Presuppositions,* ed F. Schwarz (Cham: Springer International Publishing), 215–240. doi: 10.1007/978-3-319-07980-6_10

Sanford, A. J., Moxey, L. M., and Paterson, K. B. (1996). Attentional focusing with quantifiers in production and comprehension. *Mem. Cogn.* 24, 144–155. doi: 10.3758/BF03200877

Skordos, D., and Papafragou, A. (2016). Children's derivation of scalar implicatures: alternatives and relevance. *Cognition* 153, 6–18. doi: 10.1016/j.cognition.2016.04.006

Sperber, D., and Wilson, D. (1986). *Relevance: Communication and Cognition.* Cambridge, MA: Harvard University Press.

Spotorno, N., Cheylus, A., Van Der Henst, J.-B., and Noveck, I. A. (2013). What's behind a P600? Integration operations during irony processing. *PLoS ONE* 8:e66839. doi: 10.1371/journal.pone.0066839

Spychalska, M., Kontinen, J., Noveck, I., Reimer, L., and Werning, M. (2019). When numbers are not exact: ambiguity and prediction in the processing of sentences with bare numerals. *J. Exp. Psychol. Learn. Mem. Cogn.* 45:1177. doi: 10.1037/xlm0000644

Spychalska, M., Kontinen, J., and Werning, M. (2016). Investigating scalar implicatures in a truth-value judgement task: evidence from event-related brain potentials. *Lang. Cogn. Neurosci.* 31, 817–840. doi: 10.1080/23273798.2016.1161806

Szymanik, J., and Zajenkowski, M. (2013). "Monotonicity has only a relative effect on the complexity of quantifier verification," in *Proceedings of the 19th Amsterdam Colloquium*. ILLC: University of Amsterdam.

Tomlinson, J. M., Bailey, T. M., and Bott, L. (2013). Possibly all of that and then some: scalar implicatures are understood in two steps. *J. Mem. Lang.* 69, 18–35. doi: 10.1016/j.jml.2013.02.003

Van Tiel, B., and Kissine, M. (2018). Quantity-based reasoning in the broader autism phenotype: a web-based study. *Appl. Psycholinguist.* 39, 1373–1403. doi: 10.1017/S014271641800036X

van Tiel, B., and Pankratz, E. (2021). Adjectival polarity and the processing of scalar inferences. *Glossa J. Gen. Linguist.* 6:32. doi: 10.5334/gjgl.1457

van Tiel, B., Pankratz, E., and Sun, C. (2019). Scales and scalarity: processing scalar inferences. *J. Mem. Lang.* 105, 93–107. doi: 10.1016/j.jml.2018.12.002

van Tiel, B., and Schaeken, W. (2017). Processing conversational implicatures: alternatives and counterfactual reasoning. *Cogn. Sci.* 41, 1119–1154. doi: 10.1111/cogs.12362

Van Tiel, B., Van Miltenburg, E., Zevakhina, N., and Geurts, B. (2014). Scalar diversity. *J. Semant.* 33, 137–175. doi: 10.1093/jos/ffu017

Wason, P. C. (1959). The processing of positive and negative information. *Q. J. Exp. Psychol.* 11, 92–107. doi: 10.1080/17470215908416296

Wason, P. C. (1961). Response to affirmative and negative binary statements. *Br. J. Psychol.* 52, 133–142. doi: 10.1111/j.2044-8295.1961.tb00775.x

Weiland, H., Bambini, V., and Schumacher, P. B. (2014). The role of literal meaning in figurative language comprehension: evidence from masked priming ERP. *Front. Hum. Neurosci.* 8:583. doi: 10.3389/fnhum.2014.00583

Wilson, D., and Sperber, D. (1998). "Pragmatics and time," in *Relevance Theory: Applications and Implications*, eds. R. Carston and S. Uchida (Amsterdam: John Benjamins), 1–22. doi: 10.1075/pbns.37.03wil

Wilson, D., and Sperber, D. (2012). *Meaning and Relevance.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139028370

Zhang, J., and Wu, Y. (2022). Epistemic reasoning in pragmatic inferencing by non-native speakers: the case of scalar implicatures. *Second Lang. Res.* doi: 10.1177/02676583211069735