A closer look at the sources of variability in scalar implicature derivation: a review

Ahmed Khorsheed[1]* and Nicole Gotzner[2]*

[1]Faculty of Modern Languages and Communication, Universiti Putra Malaysia, Serdang, Malaysia,
[2]Institute of Cognitive Science, Osnabrück University, Osnabrück, Germany

For more than 20 years, studies in experimental pragmatics have provided invaluable insights into the cognitive processes involved in deriving scalar implicatures and achieving inferential comprehension. However, the reports have always contained a notable degree of variability that remained inadequately discussed in the literature. For instance, upon closer inspection of the experimental record, one can always find a group of individuals who tend to be largely pragmatic, overwhelmingly logical, or sometimes mixed not showing a clear preference. There also exist newly-devised paradigms that prompt a radically different type of response than other paradigms, and thus new evidence casting doubt on long-established findings in the field. More recent research on scalar diversity further suggests that differences in the semantic structure of scalar words can lead to differing rates of scalar implicatures and can modulate the time invested in pragmatic processing. Indeed, one can contend that the current empirical landscape on scalar implicatures can be characterized as having three primary sources of variability: inter-individual, methodological, and linguistic. What factor or factors are behind these patterns of variability, and how can we interpret them in light of a pragmatic theory? This paper has a 2-fold objective: one is to review the previous experimental record on scalar implicatures from variability-based lenses, and the other to discuss the factor(s) that could account for this observed variability in the literature. Avenues for future research are provided.

KEYWORDS

scalar implicature, semantics, pragmatics, variability, testing paradigms, scales

1. Introduction

For over two decades, the study of scalar implicatures has served as a testing bed for several accounts in experimental pragmatics. Studies mainly investigated how readers and listeners would evaluate utterances containing weak scalar words (e.g., <some, all>, <or, and>, <might, must>) and how the implicature embedded in them would manifest itself in real-time processing, especially if it is generated automatically as posited by the default account (e.g., Levinson, 2000), or with processing costs by Relevance Theory (e.g., Sperber and Wilson, 1986; Wilson and Sperber, 1998). For this theoretical debate, a lot of experimentation has been carried out. People used a variety of testing techniques which include but are not limited to eye-tracking paradigms (e.g., Huang and Snedeker, 2009a, 2018; Politzer-Ahles and Matthew Husband, 2018), mouse-tracking paradigms (e.g., Tomlinson et al., 2013), event-related potentials (ERPs) (e.g., Noveck and Posada, 2003; Chevallier et al., 2010; Nieuwland et al., 2010; Hunt et al., 2013; Spychalska et al., 2016), sentence verification tasks (e.g., Bott and Noveck, 2004; Chevallier et al., 2008; Bott et al., 2012), dual tasks (e.g., De Neys and Schaeken, 2007; Marty and Chemla, 2013),

and reading comprehension vignettes (Breheny et al., 2006), among others. Results generally concluded in favor of Relevance Theory (see Noveck, 2018; Breheny, 2019; Khorsheed et al., 2022b for reviews). This evidence was also replicated across different populations and languages (Katsos et al., 2016), and it was obtained from materials covering a wide range of scalar terms (van Tiel et al., 2016). However, upon closer inspection of the data in these studies and relevant literature, one finds different variability patterns that are worth attending to. In our review, we have identified three specific categories of variability: inter-individual, methodological, and linguistic.

Regarding inter-individual variability, numerous studies have consistently revealed that individuals always prefer one particular reading over another. For instance, in a study conducted by Noveck and Posada (2003), which used the "Some cats have ears" type of material, participants were divided into two qualitatively distinct groups: those who primarily adhere to a logical reading, and those who predominantly adopt a pragmatic interpretation. Sometimes there existed a smattering of "mixed" participants who do not seem to be consistent in preferring a certain reading (see Bott and Noveck, 2004). This observation was also witnessed in other studies mainly employing sentence verification paradigms to measure participants' response rates (see Feeney et al., 2004; Breheny et al., 2006; De Neys and Schaeken, 2007; Pouscoulous et al., 2007; Bott et al., 2012; Heyman and Schaeken, 2015; Antoniou et al., 2016; Mazzaggio and Surian, 2018). However, the precise factors contributing to this inter-individual variability are still unknown.

When it comes to methodology, recent reports have highlighted that certain testing paradigms elicit a completely different type of response than other paradigms. For instance, in the context of binary judgment tasks, the developmental literature has shown that small children, even linguistically competent, are less sensitive to underinformative *some* expressions compared to adults (e.g., Noveck, 2001; Papafragou and Musolino, 2003; Pouscoulous et al., 2007). However, when modifications are made to the binary judgment task, such as introducing a middle response option (Katsos and Bishop, 2011), incorporating a rating scale (Jasbi et al., 2019), or employing a rewarding system (Bleotu et al., 2021), children demonstrate greater sensitivity to implicatures than previously observed. These methodological adjustments and their resulting effects were also observed in the adult literature (e.g., Grodner et al., 2010; Politzer-Ahles and Fiorentino, 2013). Such research has revealed that adults can process scalar implicatures as fast as logical responses, thus providing evidence opposing the bulk of findings in the literature.

Another intriguing aspect emerging from the literature, especially from work on semantic-pragmatic interaction enterprises, is the influence of different linguistic structures on the derivation of scalar implicatures (Verstraete, 2005; Baker et al., 2009; van Tiel et al., 2016, 2019; Gotzner et al., 2018; Sun et al., 2018; van Tiel and Pankratz, 2021). For instance, van Tiel et al. (2016) assessed the triggering phenomenon for 43 weak and strong scalar pairs that come from a variety of grammatical categories, including adjectives, auxiliary verbs, main verbs, and adverbs. Results showed that the proportion of endorsement rates to these tested scalar terms is highly variable: a crucial factor accounting for this variance pertains to whether the strong scalar

expression denotes an upper-bound, or due to the nature of the underlying measurement scales these expressions participate in (see also Gotzner et al., 2018).

This line of research currently casts doubt on the assumption that a single mechanism accounts for all scalar implicatures and suggests that variability could either stem from different processing paths associated with the inference itself or from how alternatives are composed for different scalar expressions.

In a nutshell, with this quick panorama presented above, one can say that the empirical landscape on scalar implicatures has three distinct patterns of variability that deserve attention and further investigation: inter-individual variability, variability triggered by differences in the apparatus of testing paradigms (i.e., methodological), and variability engendered by differences in the semantic structure of lexical scales (i.e., linguistic). What factor or factors are behind this variability? This paper has a 2-fold objective: one is to review the previous experimental record on scalar implicatures, and the other to discuss the factors that underlie these variability patterns. In what follows, we will discuss each category of variability in a separate section and showcase the underlying factors with relevant discussions and directions for future research.

2. Inter-individual variability

Recent reports in the literature show that participants vary in the tendency with which they derive scalar implicatures. While some individuals consistently prefer logical readings, others pragmatic (e.g., Noveck and Posada, 2003; Bott and Noveck, 2004; De Neys and Schaeken, 2007). One may ask why this occurs? This question has recently been the center of numerous discussions in the literature, and general results suggest that several factors contribute to this phenomenon. These factors include, but are not limited to, individual differences in working memory capacity (Feeney et al., 2004; Dieussaert et al., 2011; Marty and Chemla, 2013; Antoniou et al., 2016), Theory of Mind ability (Fairchild and Papafragou, 2021; Khorsheed et al., 2022a), and other personality characteristics which may include one's social skill and/or communication skill (Nieuwland et al., 2010; Yang et al., 2018), systemizing skill (Pijnacker et al., 2009; Chevallier et al., 2010; Barbet and Thierry, 2016), language proficiency (Antoniou et al., 2020; Khorsheed et al., 2022a), and attitudes toward honesty and integrity (Feeney and Bonnefon, 2013; Mazzarella, 2015), among others. In our view, these variability factors can be forced into two distinct groups of factors: *internal* cognitive factors and *external* social factors. While the internal factors are essentially related to the involvement of internal cognitive processes in scalar implicature derivation (i.e., working memory capacity and theory of mind), the external social factors pertain to personality traits and characteristics that may impact an individual's decision to accept or reject the scalar implicature (e.g., age, proficiency, politeness). The discussion below examines these two qualitatively-distinct groups of factors and their effects on scalar implicature derivation.

2.1. Internal cognitive factors

2.1.1. Working memory capacity

Many studies agree that scalar implicature derivation is a process that draws on one's working memory resources. This observation was mainly supported by experiments utilizing response deadlines (Bott and Noveck, 2004; Bott et al., 2012), dual tasks (De Neys and Schaeken, 2007; Dieussaert et al., 2011; Marty and Chemla, 2013), and direct measures of working memory capacity (Antoniou et al., 2016; Fairchild and Papafragou, 2021). For instance, Bott and Noveck (2004, Experiment 4) tested the likelihood that scalar implicature computation involves cognitive resources by manipulating the time available to participants i.e., *Long* condition vs. *Short* condition. While the former condition allows participants to have a relatively long time to respond (3 s), the latter allows only a relatively short time (900 ms). Crucially, the latter condition was designed to pressure participants and narrow down their cognitive resources. As Bott and Noveck (2004) pointed out, participants with more cognitive resources (i.e., in the *Long* condition) would draw scalar implicatures more often than participants with curtailed cognitive resources (i.e., *Short* condition). The task required French participants to read categorical sentences (e.g., "Some elephants are mammals") and judge if they are true or false among other control sentences. Bott and Noveck (2004) found that participants were more successful at interpreting the scalar implicature when given more time to draw on their working memory resources, but they were less likely to draw the implicature when their cognitive resources were rendered limited. Notably, this effect was only observed in the underinformative items but not in the patently-true or patently-false control sentences (for similar evidence, see De Neys and Schaeken, 2007; Bott et al., 2012; Marty and Chemla, 2013; Tomlinson et al., 2013). At the time, this finding had implications for studies interested in inter-individual variability. In other words, the propensity to be dominantly logical or pragmatic in a given task might be a proxy of individual differences in working memory capacity: those with greater working memory capacity may derive more scalar implicatures than participants with low working memory capacity (see Feeney et al., 2004; Banga et al., 2009; Dieussaert et al., 2011; Janssens et al., 2014; Heyman and Schaeken, 2015; Antoniou et al., 2016).

For instance, in a direct attempt to investigate how individual differences in working memory capacity may influence the derivation rate of scalar implicatures, Antoniou et al. (2016) conducted a study that employed two measures of working memory alongside a pragmatic task. While the pragmatic task involved statements such as *There are hearts on some of the cards* that participants have to judge as true or false based on a visual display showing hearts on all five cards, the working memory measures comprised a backward digit span task and reading span task. In a regression model, Antoniou et al. (2016) tested the relationship between participants' scalar implicature derivation rate and their overall working memory scores, in addition to a battery of personality measures included in the model. Their results revealed that only working memory capacity can account for the variance in participants' scalar implicature derivation rate. Specifically, individuals with greater working memory capacity were more

likely to reject underinformative sentences compared to those with lower working memory capacity. Their finding was replicated in some studies (e.g., Yang et al., 2018; Fairchild and Papafragou, 2021), but not observed elsewhere (Banga et al., 2009; Dieussaert et al., 2011). For instance, Banga et al. (2009) investigated the relationship between working memory capacity and scalar implicature derivation in sentences such as *Some elephants have trunks* and their findings did not reveal any significant difference in the rate of scalar implicatures between individuals with lower and higher working memory abilities (see also Janssens et al., 2014; Heyman and Schaeken, 2015). This discrepancy currently prompts inquiries regarding the potential factors that contribute to the presence or absence of working memory effect. Could it be attributed to a confounding variable?

Antoniou et al. (2016) suggest that, besides the cognitive effort required for the calculation of the implicature, some experimental designs place more cognitive demands on participants' cognitive resources than other designs, and thus is a potential explanation for these observed discrepancies in the literature. In a similar vein, Heyman and Schaeken (2015) propose that the cognitive cost invested in scalar implicature derivation may be relatively small, not placing enough cognitive demands on the working memory resources, and thus participants with limited working memory capacity can still derive pragmatic interpretations in comparable proportions to those with high working memory capacity. Indeed, these two views seem to be corroborated by the experimental evidence obtained from dual tasks (e.g., De Neys and Schaeken, 2007; Dieussaert et al., 2011; Marty and Chemla, 2013). For instance, Dieussaert et al. (2011) used a dual-task methodology in which adult participants were instructed to evaluate the truth value of underinformative statements like *Some tulips are flowers* while their executive working memory was experimentally burdened by concurrent memorization of simple and complex dot patterns: low vs. high cognitive load. Interestingly, Dieussaert et al. (2011) found no *direct* relationship between participants' memory scores and the proportion of pragmatic responses, but they observed an interaction effect between working memory capacity and the cognitive load imposed by the memorization task. More specifically, Dieussaert et al. (2011) showed that *low* working memory capacity *itself* does not lead to fewer pragmatic interpretations unless an additional cognitive load is imposed.

While these aforementioned explanations for the working memory discrepancy seem plausible from a methodological standpoint, we contend that the presence or absence of the working memory effect in scalar implicature derivation may reflect the *varying effort* that participants would expend to make the linguistic utterance maximally relevant to their expectations. This viewpoint is consistent with Relevance Theory (e.g., Sperber and Wilson, 1986), which asserts that human cognition is geared to the maximization of relevance in communication and that the processing of an utterance and the mental effort associated with it may greatly vary by the extent to which an individual is willing to bridge the gap between the linguistic meaning of an utterance and the intended meaning of the speaker. This bridging process may involve doing enrichments, revisions, and re-organizations of existing beliefs and plans. As such, the processing cost observed in computing scalar implicatures and the demands imposed on

working memory may vary as a function of the extent to which an individual will engage in deeper mind-reading activity, or Theory of Mind (ToM) (Noveck, 2018; Fairchild and Papafragou, 2021; Ronderos and Noveck, 2023). To make this observation more evident, the following subsection will discuss ToM in relation to working memory capacity and scalar implicature derivation.

2.1.2. Theory of mind

The post-Gricean view on how scalar implicatures are made and entertained suggests that a Theory of Mind (ToM) component is integrated into the process responsible for inference-making (Noveck and Sperber, 2007; Noveck, 2018). In essence, ToM refers to one's skill to attribute beliefs, intents, and desires to others and use these attributions to make predictions about another's behavior (e.g., Apperly et al., 2010; Apperly, 2012; Bergen and Grodner, 2012). Recent research in experimental pragmatics revealed a reliable relationship between ToM and pragmatic interpretations (Marocchini and Domaneschi, 2022), including scalar implicatures (Fairchild and Papafragou, 2021; Ronderos and Noveck, 2023). However, a currently debated question is whether the involvement in ToM reasoning is part of the behavioral variation in scalar implicature derivation between individuals. An indirect response to this question comes from Antoniou et al. (2016) whose results revealed a negative relationship between participants' age and the rate of pragmatic responses. Antoniou et al. sought to explain this potential link between age and the derivation rate of scalar implicatures based on speculation suggesting that older adults are less likely to employ ToM reasoning compared to younger adults (see also Bernstein et al., 2011; Henry et al., 2013), and hence they generate scalar implicatures to a lesser extent than younger peers. However, this explanation was only speculative and lacking empirical support. Antoniou et al. (2016) suggested that one can only make certain that a conclusive link between ToM and derivation rate exists if one employs a direct measure of ToM, such as the tasks utilized in previous studies on ToM (Keysar et al., 2000, 2003; Apperly et al., 2010).

In capitalizing on Antoniou et al.'s (2016) results, a recent study by Fairchild and Papafragou (2021) investigated the role of ToM in the observed variation in scalar implicature derivation. Recognizing that ToM reasoning requires cognitive effort (Epley et al., 2004; Apperly et al., 2010; Lin et al., 2010), the authors took the executive function (EF), especially the working memory capacity, as a caveat measure in their design. Their study employed five tasks: dual scalar implicature task, auditory backward digit span task, simple scalar implicature task, and the mind in the eyes and strange stories task (see their first experiment). The composite score obtained from the digit span task and high cognitive load trials embedded in the dual scalar implicature task was taken as a measure of EF, whereas the data obtained from the mind in the eyes and strange stories tasks were used to create a composite ToM score. Fairchild and Papafragou (2021) conducted relationship analyses and their preliminary results revealed that both EF and ToM are significant predictors of scalar implicature derivation rate: participants with better EF and ToM abilities are more inclined to adopt a pragmatic interpretation of sentences such as *some dogs are mammals*. Interestingly, however, when the shared variance

between EF and ToM was controlled for, only ToM exhibited a unique contribution to the variability in pragmatic responses, that is, participants with higher ToM abilities tend to derive more scalar implicatures compared to those with lower ToM abilities (see also Khorsheed et al., 2022a for similar evidence). Currently, these results raise skepticism regarding the *direct* role of working memory in scalar implicature derivation. The presence or absence of working memory effect (or processing cost) in scalar implicature computation seems to be contingent on the level of ToM that an individual will employ to discern a speaker's intention(s) (see Ronderos and Noveck, 2023, for a more lucid account). That said, the involvement of ToM in scalar implicature derivation is possibly both *individual-* and *situation-*particularized and further investigation into these variability factors in scalar implicature processing may add valuable insights to this line of research.

2.2. External social factors

As previously discussed, the external social factors cover various personality traits and characteristics that may partly impact the tendency with which one would penalize pragmatic violations (Nieuwland et al., 2010; Feeney and Bonnefon, 2013; Mazzarella et al., 2018; Yang et al., 2018; Terkourafi et al., 2020; Khorsheed et al., 2022a). Nieuwland et al. (2010) were the first to provide breakthrough evidence falling into this category. Their work examined the impact of processing underinformative statements (e.g., *Some people have lungs*) on the N400 event-related potential (ERP) and if individual variation in N400 modulation is a product of differences in personality-related skills. Taken as a caveat, Nieuwland et al. (2010) obtained a measure of the pragmatic skill of participants through the Communication subscale in the Autism Spectrum Quotient (Baron-Cohen et al., 2001) and found that participants with a low score on the Communication sub-scale (referred to as "pragmatically skilled participants") are more sensitive to pragmatic violations than participants with a high score. The same effect was replicated by Zhao et al. (2015) who found that acoustically-presented underinformative sentences (e.g., "Some tigers have tails") elicit a prominent ERP effect in participants with a low score on the Communication and Social Skill sub-scales ("the high pragmatic ability group").

In a similar vein, Barbet and Thierry (2016) recently showed that individuals with high-systemizing skills, as measured by the Systemizing Quotient (SQ-R) (Wheelwright et al., 2006), reject pragmatic violations more frequently than those with low systemizing skills. The systemizing skill refers to the extent to which one can analyze systems, extract controlling rules, and predict outputs (Baron-Cohen et al., 2003; Wheelwright et al., 2006). According to Baron-Cohen et al. (2003), individuals with high systemizing-brain are more likely to attend to details and features in objects and systems, and they treat these particular details as measured variables. They can also identify the effect of a certain operating on the input by finding its effects elsewhere in the output (correlation rules): "If the speaker says X, then A (input) changes to B (output), and thus a strong sensitivity to patterns". Given this, some hearers may base their judgments on statistical patterns that help them gauge the likelihood that

a potential interpretation is relevant to the speaker's intended meaning. This view was supported by work on individuals with high-functioning autism and Asperger's syndrome (Pijnacker et al., 2009; Chevallier et al., 2010). Results showed that participants with high systemizing abilities are more likely to reject sentences with pragmatic violations. This finding was also tested in other contexts among neurotypical individuals where results revealed a positive relationship between one's high systemizing ability and their propensity to reject underinformative sentences (Barbet and Thierry, 2016).

According to several reports (Feeney and Bonnefon, 2013; Heyman and Schaeken, 2015; Antoniou et al., 2016; Mazzaggio and Surian, 2018), the personality-based account covers a broad range of factors that extend beyond those previously discussed. These include, but are not limited to, extroversion, agreeableness, consciousness, neuroticism, and openness as measured by the Big Five Inventory (B5) (John et al., 2008), communication, attention control, attention switching, attention to detail, and imagination by the Autism Spectrum Quotient (ASQ) (Baron-Cohen et al., 2001), and politeness and honesty as measured by the Honesty/Integrity/Authenticity scale (HIA) (Goldberg et al., 2006), among others. Nevertheless, it is noteworthy that while certain autistic traits seem to exhibit a significant relationship with the derivation rate of scalar implicatures, other general personality characteristics do not show such a link (see Heyman and Schaeken, 2015; Antoniou et al., 2016; Khorsheed et al., 2022a). This discrepancy could potentially be driven by theoretical considerations: while the aforementioned measures obtained from the ASQ and SQ-R are indicative of autistic-related traits whose *pronounced* prevalence among some individuals may interact with the internal cognitive processes responsible for scalar implicature computation (see Noveck, 2018, p. 184–192), the effects obtained from the B5, HIA, or other general personality characteristics are possibly contingent on situational factors (e.g., Feeney and Bonnefon, 2013; Mazzarella, 2015; Holtgraves and Perdew, 2016; Terkourafi et al., 2020), or may be linked to metalinguistic and egocentric decisions. For instance, an individual may accept an underinformative statement even when they realize that, for example, the use of *some* in "Some cats have ears" is underinformative. We suggest that this variability is arbitrary and not consistent; and therefore, its impact on the derivation rate of scalar implicatures is not certain.

3. Methodological variability

As spelled out extensively by Noveck and Sperber (2007), experimental work needs to ensure that a given effect is robust. That is, one wants to replicate the same result over and over again across a variety of comparable tasks and testing paradigms. When two studies produce comparable outcomes, this bolsters each other's findings. However, when a study produces a new kind of result in a predictable manner, it pays to tease out the factors that underlie this observed effect.

In fact, for more than two decades now, studies on scalar implicatures have used a wide variety of techniques to examine whether the process responsible for scalar implicature derivation is cognitively costly or cost-free. The initial data discovery obtained

from the developmental literature (Noveck, 2001) and adult literature (Bott and Noveck, 2004) showed that scalar implicatures involve processing costs. However, at the time, this evidence was considered counter-intuitive and one worry was that these results seemed to be based on "reasoning" tasks, and thus risked being not generalizable to other tasks such as sentence processing. Apropos, this has led scholars to design diverse testing paradigms with varying degrees of veracity. In the context of adult processing, this led to investigating scalar implicatures using text comprehension vignettes (Breheny et al., 2006), sentence processing via eye-tracking paradigms (Huang and Snedeker, 2009a), and dual tasks to show that an added cognitive burden can impair pragmatic processing (De Neys and Schaeken, 2007). These further follow-up studies led to further confirmation of the initial finding that scalar implicatures involve processing costs (e.g., Dieussaert et al., 2011; Marty and Chemla, 2013; Tomlinson et al., 2013; Spychalska et al., 2016; van Tiel and Schaeken, 2017; Fairchild and Papafragou, 2021).

However, with time in the last decade, the adult literature witnessed several experimental and contextual manipulations in which no processing costs are observed. This trend was notably seen in the work of Grodner et al. (2010) (see also Bergen and Grodner, 2012; Politzer-Ahles and Fiorentino, 2013; Hartshorne et al., 2015; Barbet and Thierry, 2018). For instance, Grodner et al. (2010) proposed that if the partitive "Some of" is phonetically reduced to "summa," if the paradigm does not involve numbers within the subitizing range, and if the context draws attention to the underinformative use of "summa" in contexts where "alla" (meaning "all of") is the case, then "summa" cases may appear to be as reactive as "alla" cases, although not entirely (see their first-half "summa" results on p. 46). In a similar vein, Barbet and Thierry (2018) investigated scalar implicature on single words, including critical quantified terms such as "some" by utilizing a Stroop task that was contextually neutral and not biased toward either an upper-bound or lower-bound reading. Their study aimed to explain discrepancies observed in previous context-dependent reading experiments, especially the processing effort observed in reading the "some-region" in Breheny et al. (2006) and Bergen and Grodner (2012), but not in Politzer-Ahles and Fiorentino (2013), or Hartshorne et al. (2015). Barbet and Thierry (2018) did not find an N450 effect associated with pragmatic interpretation, they did observe a P600 effect potentially indicative of pragmatic processing (Spotorno et al., 2013; Spychalska et al., 2016).

In the context of children and implicature understanding, much work has been devoted to accounting for Noveck's original discovery that younger children tend to be largely logical before gradually incorporating pragmatic considerations as they grow older. Specifically, Noveck (2001) investigated how 5-, 7-, and 9-year-olds would evaluate utterances with weak logical terms (e.g., X *might* be Y) in contexts in which the stronger alternative (X *must* be Y) is true. Noveck's results showed that 5-year-old children, even linguistically competent, are more logical than other older children and adults in responding (i.e., treating *might* as compatible with *must* but not exclusive). As Noveck (2001) pointed out, children's ability to provide more nuanced interpretations of underinformative utterances develops progressively with age. Noveck's finding sparked considerable interest among researchers, leading them to explore whether children's limited engagement in pragmatic computations compared to adults represents a genuine

developmental effect or is influenced by other experimental factors. This in itself prompted scholars to develop more user-friendly paradigms. While for some researchers the interest was to prove that younger children can perform in an adult-like manner (Papafragou and Musolino, 2003), for others it was to demonstrate that a developmental effect still exists, even with simplified tasks (Pouscoulous et al., 2007).

For instance, Papafragou and Musolino (2003) conducted a study in Greece in which 5-year-old children and adults were presented with a set of vignettes narrated by an experimenter using toy-props. One of these vignettes described three horses jumping over a log, followed by a summary provided by a puppet stating that "Some of the horses jumped over the log." Papafragou and Musolino (2003) showed that an overwhelming majority of the children (88%) perceived the puppet's response as an accurate summary of the story, but the adults did not (i.e., akin to Noveck's main finding). However, it was proposed that children's apparent logical behavior could be a result of their limited awareness of the ultimate goal of the task, which is to impose severe penalties on sentences that are not sufficiently informative. As such, in a follow-up experiment, Papafragou and Musolino (2003) made adjustments to the design and used more explicit instructions, including giving the children four warm-up lessons and training to detect the pragmatic terms anomaly. The results of their experiment revealed that while children's ability to discern underinformative statements had indeed progressed, their overall performance still fell notably short of that demonstrated by adults (∼50 vs. 90%). The training effects also seemed to dissipate when the same participants are tested a week later (see Guasti et al., 2005).

In a more recent examination of the developmental effect identified by Noveck (2001), researchers proposed that children's non-adult behavior may be linked to artifacts inherent in the binary judgment task. It was suggested that children are highly sensitive to pragmatic violations but struggle to express their pragmatic abilities when tasked with providing truthful judgments (Katsos and Bishop, 2011). To investigate this proposal, Katsos and Bishop (2011) devised a new task in which participants can rate the appropriateness of a speaker's utterance based on a 3-point scale, e.g., children can reward a puppet's underinformative response with a "small," "big," or a "huge strawberry," depending on "how good" it describes a given story. In one of their stories, Katsos and Bishop (2011) had 5-to-6-year-old children see a pile of carrots (five carrots) on the left side of a screen and a mouse carrying each one of them to the right side of the screen. The mouse consistently traverses from the right side of the screen to the pile of carrots to carry each one of them back to its initial position. At the end of the narrative, the experimenters used an animated fictional character to describe the scene by using a less-than-optimally informative statement, stating "The mouse picked up some of the carrots". According to Katsos and Bishop (2011), children would exhibit sensitivity to pragmatic violations if their evaluation of the speaker's statement merits a "medium-sized strawberry" as opposed to a "huge strawberry" (indicating a response that is somewhat forgiving but still more stringent than complete agreement). The results revealed that children almost always assigned the speaker with a "medium-sized strawberry" on the scale, indicating a heightened sensitivity to underinformative sentences compared to previous findings using binary judgment

tasks. This effect has also been replicated in other studies (Schaeken et al., 2018; Jasbi et al., 2019; Bleotu et al., 2021).

Given these findings above, a question may arise as to why children appear to be overwhelmingly "pragmatic" in rating tasks but logical in binary judgment tasks. In other words, which task provides a more valid measure of children's genuine inferential ability? Several scholars have sought to address this question. For instance, Noveck (2018) suggests that ternary judgment tasks, or tasks using rewards and ratings, are *metalinguistic* tasks that, in their essence, ask participants to focus on the utterance's "appropriateness" and/or "how well" the speaker says it, and thus risk being able to provide a reliable measure of how children *process* underinformative utterances (see also Noveck, 2018 for more alternative explanations, p. 89–90). In contrast, binary judgment tasks are deemed reasoning-based tasks that establish themselves on the assumption that listeners should be able to recognize, without explicit instruction, that "the purpose of the task is not to determine whether a given sentence is true or false in a given context but rather whether the sentence in question can be used felicitously in that context" (Papafragou and Musolino, 2003, p. 269). Therefore, children's difficulty to derive scalar implicatures in eye tracking experiments (Huang and Snedeker, 2009b), and in contexts in which the communicative expectation is made maximally relevant (Papafragou and Musolino, 2003; Pouscoulous et al., 2007), lends doubt to the claim that the developmental effect is an artifact of an overt judgment response.

Recent priming studies on scalar implicatures suggest that these discrepancies may be attributed to scale-activating confounds that these tasks may give rise to during scalar implicature derivation: while the binary judgment task does not make the relevance of the stronger alternative pronounced (e.g., Bott and Frisson, 2022), the "rewarding" task has the potential to encourage participants to take the task about an utterance's "goodness" as opposed to utterance's interpretation (Skordos and Papafragou, 2016). Alternatively, children's main problem in processing scalar implicatures may lie in their inability to retrieve and activate alternatives (Barner et al., 2011; Skordos and Papafragou, 2016; Rees and Bott, 2018; Gotzner et al., 2020), possibly due to limited processing resources (Chierchia et al., 2001; Reinhart, 2004; Barner et al., 2011). However, recent work by Jasbi et al. (2019) indicates that adults, much like children, exhibit a preference for "weak" response options in 3-, 4-, and 5-point Likert scale tasks. As such, this finding seems to challenge the notion that children's inclination to deviate from a fully logical endorsement to a "medial" response is due to the interplay between response option and activating scales, because adults, despite having the cognitive ability to call out the stronger alternative scale, still exhibit a preference for a "weak" response option in rating tasks.

In this review, our contention is that when children engage in tasks involving a scale or rewards, they perhaps assess the "fit" of a given utterance or situation based on its "numerical weight" on the scale, establishing a one-to-one correspondence. In other words, a "reward" that is smaller in size or value might be a product of a self-perceived estimation of "how appropriate" the utterance is, or "how appropriately" the speaker delivers it in a given situation (i.e., reminiscent of Noveck's (2018) explanation above). Shall our observation be true, this harks back to Degen and Tanenhaus (2015) *naturalness* account and their gumball paradigm:

participants tended to show fast reaction times and high accuracy rates when *some* is meant to depict an "intermediate" number of gumballs, but a reversed pattern when *some* is meant to describe a visualization of small set size (0 gumballs) or an unpartitioned set size (all 13 gumballs), which is considered a less natural representation of the world. This being so, the results obtained from rating tasks currently raise concerns as they may not truly reflect a genuine inference-making, but rather the case of "exact description" readings.

The other aspect of our argument is that participants, including both children and adults, may become more attracted to an "intermediate" response than a strict downright response in higher cognition tasks. Because while the optimally-false response to a scalar implicature is likely to constitute a fully-fledged inferential step (i.e., the upper limit of human processing capacity), an "intermediate" response may represent a more lenient response and less refined understanding. Notwithstanding, this proposition remains doubtful in light of the qualitative feedback furnished by children in Papafragou and Musolino's study (2003, p. 267). The majority of children's verbatim was a massive replica of the puppet's logical "some" (74%), as opposed to adults who consistently invoked the stronger "all" on the same scale (98%). Such qualitative feedback in itself suggests that children's logical responses are *characteristically* different from adults' pragmatic responses, although children and adults do act *artificially* alike when a middle response option is added to the judgment task. That said, future research should place special focus on teasing out the factors underlying the effects observed in rating tasks. The obtained results may have implications for methodological robustness in experimental pragmatics and the broader field of experimental research.

4. Linguistic variability

Recent empirical research suggests that different scalar terms possess different potentials in giving rise to implicatures (Baker et al., 2009; Doran et al., 2012; Beltrama and Xiang, 2013; van Tiel et al., 2016, 2019; Gotzner et al., 2018; Simons and Warren, 2018; Sun et al., 2018). For instance, Doran et al. (2012) examined the triggering of scalar implicatures across a range of different scale types, including gradable adjectives, ranked orderings, and quantificational items. Results showed that pragmatic interpretations are frequently less likely to arise in the case of gradable adjectives than for quantifiers, cardinal numerals, or rank orderings. Beltrama and Xiang (2013) also provided evidence that adjectival scales appear different from modal scales with respect to the implicatures they trigger. In the case of adjectival scales, it was found that weak adjectives (e.g., "decent") always trigger implicatures to the negation of the middle and strong scale-mates (e.g., "good", "excellent"), but the middle adjectives, which constitute another potential trigger on the same scale, fail to generate an upper-bounding inference; and therefore, the adjectives themselves differ in the extent to which they give rise to implicatures. Beltrama and Xiang (2013) found no such difference in modal scales, such as <possible, likely, certain>. In other words, the modal scales exhibit trigger boundaries between the weak, middle, and strong term; and therefore, the use of the weaker

term on the same scale (i.e., "possible", "likely") gives rise to the proposition that the stronger scalar term (i.e., "certain") does not hold. This adjective-modal discrepancy was explained in terms of their boundedness, specifically the unbounded nature of the adjectival measurement scales as opposed to the bounded nature of modal scales.

A more comprehensive investigation into the underlying factors contributing to the variability in scalar implicature derivation rates as influenced by different scale structures comes from work by van Tiel et al. (2016). van Tiel et al. assessed the triggering phenomenon for 43 weak and strong scalar pairs that come from a variety of grammatical categories, including adjectives, auxiliary verbs, main verbs, and adverbs. In their experiment, a fictional character made a statement involving a weak term and the participants were asked to decide whether or not this proposition implied that the statement would have been false if the expression had been replaced with strong members on the same scale. For example, participants were queried about whether the statement in (1) entailed the scalar inference in (2).

(1) John is attractive
(2) John is not stunning

The results of their study provided evidence that the proportion of endorsement rates to these tested scalar terms was highly variable. For instance, while only a few participants endorsed the potential scalar inference in (2) triggered by the weak term *attractive* in (1), almost all participants endorsed the scalar inference in sentences associated with *some*. van Tiel et al. (2016) sought to explain the factors that might be the underlying cause for this diversity in endorsement rates by examining the semantic distance between the weaker and stronger term, their association strength, the availability of the stronger term, its relative frequency, and the presence of an upper-bound on the underlying measurement scale. van Tiel et al. (2016) demonstrated that both the *upper boundedness* and *semantic distance* of scalar terms are the only significant factors that affect the rate of inferred implicatures (see also Stateva et al., 2019 for work on cross-linguistic differences relating to the numerical bounds by quantifiers).

Benz et al. (2018) later revisited van Tiel et al.'s (2016) methodology raising the concern that their task may have triggered negative strengthening. In other words, the experimental material in van Tiel et al.'s study involved scalar terms whose interpretation may lead to negating the stronger scale-mate which may in turn give rise to negative strengthening. For example, the utterance "John is not stunning," in sentence (2) above, may be strengthened to convey that "John is rather ugly", and therefore triggering a confound that is incompatible with the semantic meaning of attractive. This means that the participants may have derived the scalar implicature but they decided to cancel it since the strengthened enrichment of the stronger scale-mate appeared in conflict with their interpretation of the implicature-modified weaker term.

In a follow-up study, Gotzner et al. (2018) tested a set of 70 adjective pairs balanced in scale structure and found that the endorsements of scalar implicature are "anti-correlated with

the degree of negative strengthening of the stronger scale-mate". Similar to the findings of van Tiel et al. (2016), the upper-bound entailment and semantically distant scale-mates gave rise to higher endorsement rates in the scalar implicature task. Gotzner et al. (2018) also demonstrated that adjectives are not *per se* less likely to yield low endorsement rates for scalar implicatures but their behavior depends on the scale structure underlying semantics of scalar expressions. For example, adjectival scales like <possible, certain>, which denote a lower and an upper-bound, behave similarly to the <some, all> scale. Gotzner et al. (2018) found that polarity, the adjectival extremeness of the strong term (e.g., "gigantic" vs. "large"), and the nature of the standard invoked by the weaker scale-mate (minimum standard adjectives like *dirty* vs. relative adjectives like *large*) accounted for about 66% of the observed variance in the endorsement rates.

More recently, this scalar diversity work has been taken to the domain of language processing (van Tiel et al., 2019; van Tiel and Pankratz, 2021). For instance, van Tiel et al. (2019) examined the processing of both positive and negative scalar words and found that rejecting underinformative utterances containing *positive* scalar words (e.g., "might," "some," "or") consistently led to processing slowdowns, whereas rejecting utterances with *negative* scalar words (e.g., "low" and "scarce") did not result in noticeable processing delays. This work proposed that the processing cost observed in previous studies may not generalize to the entire family of scalar words, and that different scalar words may undergo distinct processing mechanisms (see Khorsheed et al., 2022b for review and further discussion).

In another interesting study, Alexandropoulou et al. (2022) examined how the scale structure underlying different types of adjectives may affect the derivation of scalar implicatures. They employed an incremental decision task where participants read sentences with temporal ambiguity and had to distinguish a target referent ("warm water") from a competitor ("hot water"). The visual scene either contained a contrast item (cold water) or no contrast item (along with one or two unrelated distractors). Alexandropoulou et al. (2022) found distinct verification strategies for different classes of adjectives following the role of context in their lexical semantics (building on Aparicio et al., 2016). Specifically, the immediate visual context facilitated the derivation of scalar implicatures triggered by relative adjectives ("warm but not hot"), whereas for minimum-standard adjectives scalar implicatures ("breezy but not windy") were computed robustly and independently of the visual context. The authors concluded that different kinds of scalar meaning are computed incrementally and potentially in parallel.

Overall, an important unresolved question arising from this line of research is whether strong terms are less likely to be activated or if different inference mechanisms are involved across different scalars (see also Gotzner and Romoli, 2022 for further discussion). To tease apart how alternatives are constructed across scales and how listeners derive inferences about those alternatives, future work should systematically compare different scalar expressions in the well-established processing paradigms and derive predictions for the different behavior of scalar expressions, as based on their specific lexical-semantic properties. In essence, this body of work sheds new light on the borderline between semantic and pragmatic meaning.

5. Conclusion

This paper has reviewed the experimental record on scalar implicatures and explored the factors impacting their derivation, including inter-individual, methodological, and linguistic factors. It has highlighted how this variability contributed to our understanding of the conditions, mechanisms, and factors involved in the emergence and processing of scalar implicatures, while also underscoring the challenges it poses to existing theories and scholars in the field. Currently, the empirical evidence does not conclusively favor one pragmatic account over others, leaving room for ongoing debates and discussions. In light of this, our paper provides new directions for future research that may aid in resolving key debates surrounding the processing of scalar implicatures.

Author contributions

AK prepared the first draft of the manuscript, except for Section 4 which was jointly composed with NG. Both authors have made revisions to the paper and share responsibility for all parts. Both authors contributed to the article and approved the submitted version.

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.

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

Alexandropoulou, S., Discher, H., Herb, M., and Gotzner, N. (2022). "Incremental pragmatic interpretation of gradable adjectives: the role of standards of comparison," in *Semantics and Linguistic Theory*, eds J. R. Starr, J. Kim, and B. Öney (El Colegio de Mexico and the Universidad Nacional Autónoma de México), 32, 481–497.

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

Antoniou, K., Veenstra, A., Kissine, M., and Katsos, N. (2020). How does childhood bilingualism and bi-dialectalism affect the interpretation and processing of pragmatic meanings? *Bilingualism* 23, 186–203. doi: 10.1017/S1366728918001189

Aparicio, H., Xiang, M., and Kennedy, C. (2016). "Processing gradable adjectives in context: a visual world study," in *Semantics and Linguistic Theory*, eds S. D'Antonio, M. Moroney, and C-R. Little (Palo Alto, CA: Stanford University; LSA and CLC Publications), 25, 413–432.

Apperly, I.A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *Q. J. Exp. Psychol.* 65, 825–839. doi: 10.1080/17470218.2012.676055

Apperly, I. A., Carroll, D. J., Samson, D., Humphreys, G. W., Qureshi, A., and Moffitt, G. (2010). Why are there limits on theory of mind use? Evidence from adults' ability to follow instructions from an ignorant speaker. *Q. J. Exp. Psychol.* 63, 1201–1217. doi: 10.1080/17470210903281582

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

Banga, A., Heutinck, I., Berends, S. M., and Hendriks, P. (2009). Some implicatures reveal semantic differences. *Linguist. Netherlands* 26, 1–13. doi: 10.1075/avt.26.02ban

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

Baron-Cohen, S., Richler, J., Bisarya, D., Gurunathan, N., and Wheelwright, S. (2003). The systemizing quotient: an investigation of adults with Asperger syndrome or high–functioning autism, and normal sex differences. *Philos. Transact. R. Soc. London Ser. B Biol. Sci.* 358, 361–374. doi: 10.1098/rstb.2002.1206

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., and Clubley, E. (2001). The Autism Spectrum Quotient: evidence from Asperger syndrome/high functioning autism, males and females, scientists and mathematicians. *J. Autism Dev. Disord.* 31, 5–17. doi: 10.1023/A:1005653411471

Beltrama, A., and Xiang, M. (2013). "Is 'good' better than 'excellent'? An experimental investigation on scalar implicatures and gradable adjectives," in *Proceedings of Sinn und Bedeutung 17*, eds E. Chemla, V. Homer and G. Winterstein (Paris: École normale supérieure), 81–98.

Benz, A., Bombi, C., and Gotzner, N. (2018). "Scalar diversity and negative strengthening," in *Proceedings of Sinn und Bedeutung*, eds U. Sauerland and S. Solt (Berlin: ZAS), 22: 191–203.

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

Bernstein, D. M., Thornton, W. L., and Sommerville, J. A. (2011). Theory of mind through the ages: Older and middle-aged adults exhibit more errors than do younger adults on a continuous false belief task. *Exp. Aging Res.* 37, 481–502. doi: 10.1080/0361073X.2011.619466

Bleotu, A. C., Benz, A., and Gotzner, N. (2021). Where truth and optimality part. Experiments on implicatures with epistemic adverbs. *Exp. Linguist. Mean.* 1, 47–58. doi: 10.3765/elm.1.4863

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

Breheny, R. (2019). "Language processing, relevance and questions," in *Relevance, Pragmatics and Interpretation,* eds B. Clark, K. Scott, and R. Carston (Cambridge: Cambridge University Press), 42–52.

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

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

Chevallier, C., Wilson, D., Happé, F., and Noveck, I. (2010). Scalar inferences in autism spectrum disorders. *J. Autism Dev. Disord.* 40, 1104–1117. doi: 10.1007/s10803-010-0960-8

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.

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

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

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

Doran, R., Ward, G., Larson, M., McNabb, Y., and Baker, R. E. (2012). A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language* 88, 124–154. doi: 10.1353/lan.2012.0008

Epley, N., Keysar, B., Van Boven, L., and Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *J. Pers. Soc. Psychol.* 87, 327–339. doi: 10.1037/0022-3514.87.3.327

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

Feeney, A., and Bonnefon, J. F. (2013). Politeness and honesty contribute additively to the interpretation of scalar expressions. *J. Lang. Soc. Psychol.* 32, 181–190. doi: 10.1177/0261927X12456840

Feeney, A., Scrafton, S., Duckworth, A., and Handley, S. J. (2004). The story of some: everyday pragmatic inference by children and adults. *Can. J. Exp. Psychol.* 58, 121–132. doi: 10.1037/h0085792

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., et al. (2006). The international personality item pool and the future of public-domain personality measures. *J. Res. Pers.* 40, 84–96. doi: 10.1016/j.jrp.2005.08.007

Gotzner, N., Barner, D., and Crain, S. (2020). Disjunction triggers exhaustivity implicatures in 4- to 5-year-olds: investigating the role of access to alternatives. *J. Semant.* 37, 219–245. doi: 10.1093/jos/ffz021

Gotzner, N., and Romoli, J. (2022). Meaning and alternatives. *Ann. Rev. Linguist.* 8, 213–234. doi: 10.1146/annurev-linguistics-031220-012013

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

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

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

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

Henry, J. D., Phillips, L. H., Ruffman, T., and Bailey, P. E. (2013). A meta-analytic review of age differences in theory of mind. *Psychol. Aging* 28, 826. doi: 10.1037/a0030677

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

Holtgraves, T., and Perdew, A. (2016). Politeness and the communication of uncertainty. *Cognition* 154, 1–10. doi: 10.1016/j.cognition.2016.05.005

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

Huang, Y. T., and Snedeker, J. (2018). Some inferences still take time: prosody, predictability, and the speed of scalar implicatures. *Cogn. Psychol.* 102, 105–126. doi: 10.1016/j.cogpsych.2018.01.004

Hunt, L., Politzer-Ahles, S., Gibson, L., Minai, U., and Fiorentino, R. (2013). Pragmatic inferences modulate N400 during sentence comprehension: evidence from picture-sentence verification. *Neurosci. Lett.* 534, 246–251. doi: 10.1016/j.neulet.2012.11.044

Janssens, L., Fabry, I., and Schaeken, W. (2014). 'some' effects of age, task, task content and working memory on scalar implicature processing. *Psychol. Belg.* 54, 374–388. doi: 10.5334/pb.ax

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

John, O. P., Naumann, L. P., and Soto, C. J. (2008). "Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues," in *Handbook of Personality: Theory and Research, 3rd Edn* (New York, NY: The Guilford Press), 114–158.

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

Keysar, B., Barr, D. J., Balin, J. A., and Brauner, J. S. (2000). Taking perspective in conversation: the role of mutual knowledge in comprehension. *Psychol. Sci.* 11, 32–38. doi: 10.1111/1467-9280.00211

Keysar, B., Lin, S., and Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition* 89, 25–41. doi: 10.1016/S0010-0277(03)00064-7

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

Khorsheed, A., Price, J., and van Tiel, B. (2022b). Sources of cognitive cost in scalar implicature processing: a review. *Front. Commun.* 7, 990044. doi: 10.3389/fcomm.2022.990044

Levinson, S.C. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge, MA: The MIT Press.

Lin, S., Keysar, B., and Epley, N. (2010). Reflexively mindblind: using theory of mind to interpret behavior requires effortful attention. *J. Exp. Soc. Psychol.* 46, 551–556. doi: 10.1016/j.jesp.2009.12.019

Marocchini, E., and Domaneschi, F. (2022). "Can you read my mind?" Conventionalized indirect requests and theory of Mind abilities. *J. Pragmat.* 193, 201–221. doi: 10.1016/j.pragma.2022.03.011

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., and Surian, L. (2018). A diminished propensity to compute scalar implicatures is linked to autistic traits. *Acta Linguist. Acad.* 65, 651–668. doi: 10.1556/2062.2018.65.4.4

Mazzarella, D. (2015). Politeness, relevance and scalar inferences. *J. Pragmat.* 79, 93–106. doi: 10.1016/j.pragma.2015.01.016

Mazzarella, D., Trouche, E., Mercier, H., and Noveck, I. (2018). Believing what you're told: politeness and scalar inferences. *Front. Psychol.* 9, 1–12. doi: 10.3389/fpsyg.2018.00908

Nieuwland, M. S., Ditman, T., and Kuperberg, G. R. (2010). On the incrementality of pragmatic processing: an ERP investigation of informativeness and pragmatic abilities. *J. Mem. Lang.* 63, 324–346. doi: 10.1016/j.jml.2010.06.005

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

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).

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., 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

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

Pijnacker, J., Hagoort, P., Buitelaar, J., Teunisse, J. P., and Geurts, B. (2009). Pragmatic inferences in high-functioning adults with autism and Asperger syndrome. *J. Autism Dev. Disord.* 39, 607–618. doi: 10.1007/s10803-008-0661-8

Politzer-Ahles, S., and Fiorentino, R. (2013). The realization of scalar inferences: context sensitivity without processing cost. *PLoS ONE* 8, e0063943. doi: 10.1371/journal.pone.0063943

Politzer-Ahles, S., and Matthew Husband, E. (2018). Eye movement evidence for context-sensitive derivation of scalar inferences. *Collabra Psychol.* 4, 1–13. doi: 10.1525/collabra.100

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

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

Ronderos, C. R., and Noveck, I. (2023). Slowdowns in scalar implicature processing: Isolating the intention-reading costs in the Bott & Noveck task. *Cognition.* 238, 105480. doi: 10.1016/j.cognition.2023.105480

Schaeken, W., Van Haeren, M., and Bambini, V. (2018). The understanding of scalar implicatures in children with autism spectrum disorder: dichotomized responses to violations of informativeness. *Front. Psychol.* 9, 1266. doi: 10.3389/fpsyg.2018.01266

Simons, M., and Warren, T. (2018). A closer look at strengthened readings of scalars. *Q. J. Exp. Psychol.* 71, 272–279. doi: 10.1080/17470218.2017.1314516

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

Stateva, P., Stepanov, A., Déprez, V., Dupuy, L. E., and Reboul, A. C. (2019). Cross-linguistic variation in the meaning of quantifiers: implications for pragmatic enrichment. *Front. Psychol.* 10, 957. doi: 10.3389/fpsyg.2019.00957

Sun, C., Tian, Y., and Breheny, R. (2018). A link between local enrichment and scalar diversity. *Front. Psychol.* 9, 1–12. doi: 10.3389/fpsyg.2018.02092

Terkourafi, M., Weissman, B., and Roy, J. (2020). Different scalar terms are affected by face differently. *Int. Rev. Pragmat.* 12, 1–43. doi: 10.1163/18773109-01201103

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 Pankratz, E. (2021). Adjectival polarity and the processing of scalar inferences. *Glossa* 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. (2016). Scalar diversity. *J. Semant.* 33, 137−175. doi: 10.1093/jos/ffu017

Verstraete, J. C. (2005). Scalar quantity implicatures and the interpretation of modality. problems in the deontic domain. *J. Pragmat.* 37, 1401–1418. doi: 10.1016/j.pragma.2005.02.003

Wheelwright, S., Baron-Cohen, S., Goldenfeld, N., Delaney, J., Fine, D., Smith, R., et al. (2006). Predicting autism spectrum quotient (AQ) from the systemizing quotient-revised (SQ-R) and empathy quotient (EQ). *Brain Res.* 1079, 47–56. doi: 10.1016/j.brainres.2006.01.012

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

Yang, X., Minai, U., and Fiorentino, R. (2018). Context-sensitivity and individual differences in the derivation of scalar implicature. *Front. Psychol.* 9, 1–14. doi: 10.3389/fpsyg.2018.01720

Zhao, M., Liu, T., Chen, G., and Chen, F. (2015). Are scalar implicatures automatically processed and different for each individual? A mismatch negativity (MMN) study. *Brain Res.* 1599, 137–149. doi: 10.1016/j.brainres.2014.11.049