COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 45 (2021) e12938

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12938

The Role of Executive Function and Theory of Mind in Pragmatic Computations

Sarah Fairchild,[a] Anna Papafragou[b]

[a]*Department of Psychological & Brain Sciences, University of Delaware*
[b]*Department of Linguistics, University of Pennsylvania*

Received 10 May 2018; received in revised form 18 December 2020; accepted 31 December 2020

Abstract

In sentences such as "Some dogs are mammals," the literal semantic meaning ("Some *and possibly all* dogs are mammals") conflicts with the pragmatic meaning ("*Not all* dogs are mammals," known as a *scalar implicature*). Prior work has shown that adults vary widely in the extent to which they adopt the semantic or pragmatic meaning of such utterances, yet the underlying reason for this variation is unknown. Drawing on theoretical models of scalar implicature derivation, we explore the hypothesis that the cognitive abilities of executive function (EF) and theory of mind (ToM) contribute to this observed variation. In Experiment 1, we show that individuals with better ToM are more likely to compute a scalar implicature and adopt the pragmatic meaning of an utterance; however, EF makes no unique contribution to scalar implicature comprehension after accounting for ToM. In Experiment 2, we replicate this finding and assess whether it generalizes to the comprehension of other pragmatic phenomena such as indirect requests (e.g., "It's hot in here" uttered to ask for something to be done) and metaphor (e.g., "to harvest courage"). This is the first evidence that differences in ToM are associated with pragmatic competence in neurotypical adults across distinct pragmatic phenomena.

Keywords: Pragmatics; Scalar implicature; Under-informativeness; Individual differences; Executive function; Theory of mind

1. Introduction

People do not always say exactly what they mean, and yet daily communication proceeds smoothly. According to a widely influential view of communication (Grice, 1975), this happens because listeners can process linguistic meaning at two levels: They understand the literal, *semantic* meaning of an utterance but routinely also "read between the

Correspondence should be sent to Anna Papafragou, Department of Linguistics, University of Pennsylvania, 3401-C Walnut St., Suite 300-C Wing, Philadelphia, PA 19104. E-mail: anna4@sas.upenn.edu

lines" to contextually enrich the semantic meaning with *pragmatic* inferences constrained by the speaker's intentions. These pragmatic inferences are driven by the expectation that the speaker is a cooperative interlocutor, striving to be informative, truthful, relevant, and concise—as any rational social partner would be (Grice, 1975; for an important alternative, see Sperber & Wilson, 1986). When these expectations about the speaker's communicative behavior appear to be violated, listeners are justified in attempting to reconcile the apparent violation by making an inference about the speaker's intended meaning.

Consider as an example the utterance "Some of my friends are nice." This utterance is under-informative in comparison to the more informative (and relevant) alternative "All of my friends are nice," because the speaker used the weaker term in a logical scale (*some*) when s/he could have used a stronger, more informative scalar term (*all*). In many contexts, this utterance will lead the hearer to infer that not all of the speaker's friends are nice (an inference known as *scalar implicature*; see Carston, 1995; Grice, 1975; Horn, 1972, 1984; Hirschberg, 1985; Levinson, 2000; Sperber & Wilson, 1986).

Under-informative sentences conveying encyclopedic information (e.g., "Some giraffes have long necks") have been used extensively in the literature to assess pragmatic computations because the semantic and pragmatic meanings of the sentences yield different judgments. Typically, the literal, semantic meaning of these sentences ("Some *and possibly all* giraffes have long necks") is true, while the pragmatically enriched meaning arrived at after deriving a scalar implicature ("Some *but not all* giraffes have long necks") is false. The extent to which listeners adopt the pragmatic or logical meaning of these utterances is known to vary across tasks, with some studies reporting that adults reject under-informative statements over 90% of the time (Papafragou & Musolino, 2003) and others reporting rejection rates closer to 50% (Guasti et al., 2005; Noveck, 2001). In general, adults' responses depend on task demands such as the interpretation of the implicit question that the under-informative sentence was thought to be in response to (the *Question Under Discussion*; Degen & Tanenhaus, 2015), or the predictability of the utterance given the range of available alternatives (Huang & Snedeker, 2018). Children who, unlike adults, often struggle with scalar implicatures (e.g., Noveck, 2001; Papafragou & Musolino, 2003), can also respond more pragmatically depending on the specifics of the task (Guasti et al., 2005; Katsos & Bishop, 2011; Papafragou & Musolino, 2003; Pouscoulous, Noveck, Politzer, & Bastide, 2007; Skordos & Papafragou, 2016).

Most importantly for present purposes, adults also vary in their responses to under-informative statements within a single task. Some studies have found that groups of adults have consistent response patterns, with Logical Responders accepting under-informative sentences and Pragmatic Responders rejecting them (e.g., Bott & Noveck, 2004; Noveck & Posada, 2003); other studies, however, have included an Inconsistent Responders category (Heyman & Schaeken, 2015; Politzer-Ahles, Fiorentino, Jiang, & Zhou, 2013). Assuming that individuals can be neither perfectly logical nor perfectly pragmatic, and that pragmatic responding should best be treated as a continuum (cf. Degen & Tanenhaus, 2015; Fairchild & Papafragou, 2018), an important question remains as to *why* such individual variation in pragmatic judgments emerges. A reasonable hypothesis is that stable participant characteristics (alongside task characteristics) can shift individuals' responses to scalar statements from more to less pragmatic. Here we focus on two theoretically motivated possibilities according to which individual characteristics related to executive function (EF) and theory of mind (ToM) affect pragmatic responding.[1]

1.1. Executive function and scalar implicature

Differences in EF—the collection of control processes such as working memory, inhibition, and task-switching—might be responsible for variability in adults' pragmatic judgments. According to classic models of implicature, the computation of a scalar inference from an utterance such as "Some giraffes have long necks" involves a number of steps (e.g., Geurts, 2010; Grice, 1975; Horn, 1972, 1984; Sauerland, 2004; cf. also Carston, 1995; Sperber & Wilson, 1986, for a different but still richly inferential model): Broadly speaking, the listener computes the literal meaning of the utterance ("Some and possibly all giraffes have long necks"), considers the stronger alternative to the utterance in cases where it would have been relevant to the conversation ("All giraffes have long necks"), and—assuming that the speaker is knowledgeable about giraffes—reasons that (the speaker believes that) the stronger alternative is false, otherwise the speaker would have uttered it in an effort to be informative, as required by the goals of the exchange. Eye-tracking and reading time studies indicate that processing scalar implicatures often requires additional time and is therefore cognitively costly compared to processing the literal, semantic content of the utterance (e.g., Bott, Bailey, & Grodner, 2012; Bott & Noveck, 2004; Breheny, Katsos, & Williams, 2006; Huang & Snedeker, 2009; Noveck & Posada, 2003; but see Degen & Tanenhaus, 2015; Grodner, Klein, Carbary, & Tanenhaus, 2010; Huang & Snedeker, 2018 for contexts that decrease the processing demands of scalar implicatures). An intriguing hypothesis is that individuals with more available cognitive resources—especially working memory—may be more likely to recruit these resources and calculate the implicature.

Currently, empirical evidence for the specific role of EF in scalar implicature computation is unsettled. De Neys and Schaeken (2007) found support for a role of EF—and working memory specifically—in implicature computation: In that study, participants were more likely to accept under-informative statements such as "Some dogs are mammals" when they were under higher as opposed to lower cognitive load (e.g., when they were concurrently memorizing complex as opposed to simple dot patterns; see also Dieussaert, Verkerk, Gillard, & Schaeken, 2011; Marty & Chemla, 2013). Similarly, Antoniou, Cummins, and Katsos (2016) found that working memory capacity within EF predicted the rate at which adults rejected under-informative sentences used to describe visual scenes ("There are hearts on some of the cards"). Furthermore, neuroimaging studies have demonstrated that frontal cortical regions associated with EF are activated during scalar implicature computation (Politzer-Ahles & Gwilliams, 2015; Shetreet, Chierchia, & Gaab, 2014). However, many of these studies focused specifically on EF and did not include other variables in the design (De Neys & Schaeken, 2007; Marty & Chemla, 2013), so it is possible that the observed effects of EF were due to third factors. More generally, these studies do not address the possibility that the involvement of EF in scalar implicature could be tied to the engagement of mechanisms computing speaker knowledge (see next section).

Adding to the complicated picture, a large-scale study with Dutch students by Heyman and Schaeken (2015) failed to find relationships between EF abilities and scalar implicature computation. Furthermore, a very large body of individual-differences work examining the contributions of (different aspects of) EF to pragmatic abilities in both typically and atypically developing children has failed to yield converging evidence for reliable associations between EF and specific pragmatic phenomena (see Matthews, Biney, & Abbot-Smith, 2018, for a detailed review).

1.2. Theory of mind and scalar implicature

Recall that classic models of pragmatics assume that scalar implicature computation involves the listener's assessment of the speaker's epistemic state (Carston, 1995; Geurts, 2010; Grice, 1975; Horn, 1972, 1984; Sauerland, 2004; Sperber & Wilson, 1986). Recent psycholinguistic evidence supports this assumption (Bergen & Grodner, 2012; Breheny, Ferguson, & Katsos, 2013; see also Hochstein, Bale, Fox, & Barner, 2014; Papafragou, Friedberg, & Cohen, 2018, for developmental evidence). For instance, adults are more likely to draw scalar inferences when the speaker is knowledgeable as compared to when the speaker has partial knowledge about the topic (and therefore may not know whether a stronger statement is true; Bergen & Grodner, 2012). It appears likely that ToM, the ability to compute others' mental states (Baron-Cohen, Leslie, & Frith, 1985), is implicated in the computation of scalar inferences. Given that the ability to take someone else's mental perspective varies across individuals (e.g., Apperly, 2012), it is likely that individuals with poorer ToM abilities are less likely to calculate scalar implicatures and thus might surface as logically biased responders depending on the task. However, this hypothesis has not been tested directly in adults (and even the developmental picture from individual-differences research is inconclusive; see Matthews et al., 2018).

It is worth noting that scalar implicature computation (and other pragmatic abilities—see Loukusa & Moilanen, 2009 for a review) has been investigated in adolescents and adults with autism spectrum disorder (ASD), a group that includes individuals with widely different cognitive and linguistic profiles who are generally known to have ToM deficits (Newschaffer et al., 2007).[2] In one study, ASD participants (who scored poorly on ToM measures of false belief) adopted pragmatically under-informative descriptions of story-book content more often than neurotypical controls, as anticipated by the hypothesis that ToM is involved in scalar implicature computation (Noveck, Guelminger, Georgieff, & Labruyere, 2007). Other studies have shown that ASD adolescents (Hochstein, Bale, & Barner, 2017; Pijnacker, Hagoort, Buitelaar, Teunisse, & Geurts, 2009) and adults (Chevallier, Wilson, Happé, & Noveck, 2010) are not impaired in their ability to compute implicatures. These last studies, however, either did not measure the ToM abilities of the participants (Chevallier et al., 2010; Pijnacker et al., 2009) or reported a relatively high[3] ToM score for the ASD group (Hochstein et al., 2017). Hochstein et al. (2017) argued that ToM is not required for scalar implicature. An alternative, raised by both Pijnacker et al. (2009) and Chevallier et al. (2010) about their own data, is that high-functioning individuals with autism could have basic ToM skills that are sufficient for the computation of some pragmatic inferences (see also Schaeken, Van Haeren, & Bamini, 2018).

1.3. Summary and prospectus

The evidence reviewed so far precludes firm conclusions about the potential contributions of EF and ToM to individual differences in adults' scalar implicature computation. Furthermore, the present discussion reveals several limitations in the existing literature on this topic. Perhaps the most striking limitation is that EF and ToM abilities have yet to be investigated in a single group of individuals. It is important to do so to tease apart the unique influence of each factor since ToM and EF are often correlated with one another (Apperly, 2012; Bull, Phillips, & Conway, 2008; Carlson & Moses, 2001; Carlson, Moses, & Breton, 2002; Hughes & Ensor, 2007). Relatedly, best practices in individual differences research (e.g., Cronbach, 1957; Miyake et al., 2000) require large sample sizes, multiple measures, and proof of replicability. Most of the studies reviewed above are relatively small-scale and fail to fulfill these methodological requirements (see also Matthews et al., 2018, for similar issues with developmental evidence).

A broader limitation of the studies reviewed above is that their scope does not engage the question whether individual differences documented in scalar implicature computation relate to other pragmatic phenomena. If the computation of pragmatic meaning in general involves cognitive cost that results from holding and manipulating representations in working memory (cf. Antoniou et al., 2016; De Neys & Schaeken, 2007, among others), further types of pragmatically enriched meaning should also be expected to incur similar costs and be associated with EF (working memory) abilities (see Chiappe & Chiappe, 2007; Mashal, 2013, for evidence from metaphor comprehension). Similarly, if pragmatic computation is a species of intention recognition and involves the calculation of what the speaker had in mind (Grice, 1975; Sperber & Wilson, 1986), the involvement of ToM should be fairly general across pragmatic phenomena. In support of this hypothesis, neuroimaging research has shown that cortical regions, most notably the right temporo-parietal junction (rTPJ) known to engage in ToM (Saxe & Kanwisher, 2003), are activated during the comprehension of metaphors (Prat, Mason, & Just, 2012), indirect requests (Van Ackeren, Casasanto, Bekkering, Hagoort, & Rueschemeyer, 2012), irony (Eviatar & Just, 2006), and jokes (Feng, Ye, Mao, & Yue, 2014). Furthermore, patients with lesions in these ToM areas have impairments in processing metaphors (Champagne-Lavau & Joanette, 2009) and jokes (Winner, Brownell, Happé, Blum, & Pincus, 1998). Interestingly, in some of this patient work, the relationship between ToM and pragmatic abilities holds even when EF is intact (Champagne-Lavau & Joanette, 2009). However, at present, we lack evidence connecting individual differences in EF or ToM to performance across several pragmatic phenomena.

Here we present two experiments designed to address these crucial gaps in the literature. In Experiment 1, we seek to determine the unique contributions of EF and ToM to scalar implicature, a domain known for variation in judgments and a central topic in the study of pragmatic processing and development. In Experiment 2, we investigate the influence of EF and ToM on metaphor and indirect request comprehension, in addition to scalar implicature. In these experiments we go beyond prior work that typically investigated only one factor at a time in the context of scalar implicature derivation. Furthermore, we seek to implement best practices in individual differences research by including large sample sizes (approximately 200 participants for each experiment), multiple measures (to the extent possible) of pragmatic ability, EF, and ToM, as well as evidence of replicability. Finally, we expand the empirical scope of prior work by investigating the mechanisms underlying individual variation in pragmatic computations in domains other than scalar implicature.

2. Experiment 1: Computing scalar implicatures

In Experiment 1, we investigated the role of EF and ToM on scalar implicature computation using two sources of evidence. The first source was a Dual Scalar Implicature Task inspired by De Neys and Schaeken (2007). In this task, participants judged the extent to which under-informative (and other types of) sentences made sense while holding simple (Control condition) or complex (Load condition) dot patterns in memory. The goal was to assess the contribution of EF to scalar implicature by comparing Control and Load memory conditions. According to the logic in De Neys and Schaeken (2007), under-informative statements should be judged as making more sense on Load trials as compared to Control trials, in keeping with their finding that individuals are more logical (i.e., less likely to give implicature-based responses) under heavy cognitive load. This prediction rests on the assumption that EF is recruited in order to compute a scalar implicature. According to the same line of reasoning, ratings for the control sentences, where no implicature is involved, should not differ by cognitive load.

The second source of evidence was based on individual measures of participants' pragmatic and cognitive abilities. We administered a simple scalar implicature task that assessed implicature computations with the same participants but only asked them to judge whether under-informative and fully informative statements made sense or not. This binary judgment task was modeled after a widely used paradigm in the literature on scalar implicature (e.g., Bott et al., 2012; Bott & Noveck, 2004; De Neys & Schaeken, 2007; Pijnacker et al., 2009; Slabakova, 2010). Crucially, we related performance on this task to measures of EF and ToM from the same individuals to investigate the unique contribution of each component to pragmatic inference. As in the Dual Scalar Implicature Task, our EF measure targeted working memory because of evidence linking it to higher order cognition in general (Engle, Tuholksi, Laughlin, & Conway, 1999; Kane & Engle, 2002; Kane et al., 2004) and scalar implicature in particular (Antoniou et al., 2016; De Neys & Schaeken, 2007; Marty & Chemla, 2013). Our ToM tasks targeted advanced abilities to interpret facial (eye) expressions and to explain behavior in mental-state terms, and they had previously been used with adults (Baron-Cohen, Jolliffe, Mortimore, &

Robertson, 1997; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001; Jolliffe & Baron-Cohen, 1999).

2.1. Method

2.1.1. Participants

Two hundred monolingual English speakers aged 18–47 (M = 28.68, SD = 4.84) living in the United States, 89 of whom were female, were recruited from Amazon's Mechanical Turk to participate in Experiment 1. Participants were compensated $1.50 for the 15-min study. Data from 22 individuals who reported to be bilingual and/or diagnosed with ASD were excluded, leaving 178 participants for analysis.

2.1.2. Materials and procedure

Participants completed five tasks in the order described below: dual scalar implicature task, auditory backward digit span task, simple scalar implicature task, and the mind in the eyes and strange stories tasks. Performance on the high cognitive load memory trials of the dual scalar implicature task and the data from the digit span task were used to create a composite EF score. Data from the mind in the eyes and strange stories tasks were used to create a composite ToM score.

Stimuli for the dual scalar implicature task and simple scalar implicature task can be found in the Supplementary Material.

2.1.2.1. Dual Scalar Implicature Task: At the beginning of each trial, participants were presented with a pattern of three or four dots on a 3 × 3 grid and were instructed to remember the pattern. There were two types of patterns which represented two cognitive load conditions (following Bethell-Fox & Shepard, 1988 and De Neys & Schaeken, 2007): control patterns that were simple to remember, in which there were three dots in a horizontal, vertical, or diagonal row (a "one-piece" arrangement); and Load patterns

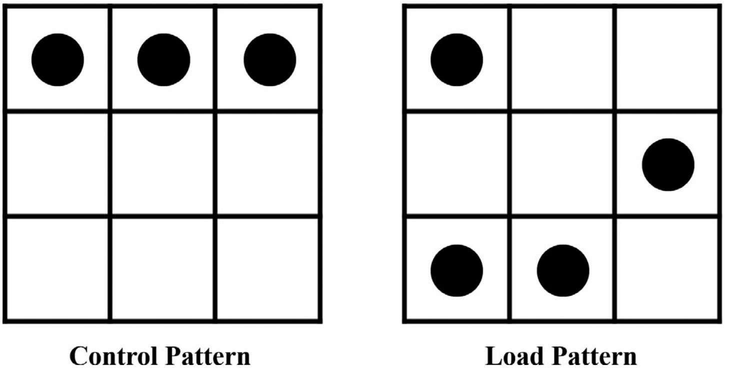


Fig. 1. Examples of control (left) and load (right) patterns used in the dual scalar implicature task in Experiment 1.

designed to increase cognitive demands, in which there were four dots arranged in a "three-piece" pattern. Examples of dot patterns are presented in Fig. 1.

On the next screen, participants read a sentence and were asked to rate it on a scale from 1 (Very Bad—Doesn't make sense) to 5 (Very Good—Makes perfect sense). Sentences of four types (see Table 1) used in our previous studies (Fairchild & Papafragou, 2018) were presented to participants: true but under-informative sentences beginning with *some* (henceforth Under-Informative), true and felicitous sentences beginning with *some* (henceforth, True [Some]), true and felicitous sentences beginning with *all* (henceforth, True [All]), and false sentences beginning with *all* (henceforth, False). The Under-Informative sentences served as the critical trials, and the other three sentence types were treated as control sentences. The four sentence types did not differ from one another in sentence length as measured in either words or syllables (all ps $> .1$). Overall, we expected that True (Some) and True (All) sentences would be rated higher than Under-Informative sentences, and those would in turn be rated higher than false sentences (see also Katsos & Bishop, 2011). Of interest was whether ratings for Under-Informative sentences would differ depending on whether participants were processing Control or Load patterns. After rating the sentence, participants were asked to recreate the dot pattern by clicking on the appropriate squares in the grid. There were 80 total trials in the task presented in a random order for each participant: 20 trials of each sentence type, half with Control patterns and half with Load patterns (counterbalanced across participants resulting in two lists).

2.1.2.2. Load trials memory in the dual scalar implicature task and auditory backward digit span task (EF measures): Participants' accuracy in recreating the dot patterns on the Load trials of the dual scalar implicature task described above served as one EF (spatial working memory) measure. An auditory backward digit span task provided an additional EF (verbal working memory) measure (Wechsler, 1944). This second task consisted of 12 trials. On each trial, participants heard a computerized voice (Apple's "Samantha") utter 5, 6, 7, or 8 digits between 1 and 9. Then participants were asked to enter the string of digits in reverse. A memory span score representing participants' EF was calculated for each participant by taking the number of digits in the longest span correctly recalled (maximum possible score of 8). This backward digit span was chosen specifically as it provides a truer measure of working memory than, for example, a forward digit span task because participants are required to perform operations on stored material (Diamond, 2013). Furthermore, this measure has good test–retest (Wechsler,

Table 1
Examples of stimuli used in the dual scalar implicature task of Experiment 1

Sentence Type	Example
Under-informative	Some baseballs are round
True (Some)	Some people have pets
True (All)	All snow is cold
False	All women are doctors

1981) and split-half (Waters & Caplan, 2003) reliability. A composite EF measure was derived by calculating z-scores from the mean number of correct Load trials and memory span on the auditory backward digit span task for each participant and then taking the sum of the two z-scores.

2.1.2.3. Simple scalar implicature task: Ten sentences borrowed from Bott and Noveck (2004) were presented individually in a random order, and participants judged whether they were "Good" or "Bad." Participants were instructed that a Good sentence is one that made sense, and a Bad sentence is one that did not make sense. Five of the sentences were Under-Informative ("Some dogs are mammals") and five were Informative ("Some fish are tuna"). The Informative sentences served as control trials and were designed to be consistently judged as "Good." The critical Under-Informative sentences could also be judged as "Good" if the participant adopted the literal interpretation of the statement (e.g., "Some *and possibly all* dogs are mammals"), or they could be judged as "Bad" if the participants derived a scalar implicature (e.g., "*Not all* dogs are mammals"). Results from this task were related to EF and ToM measures.

2.1.2.4. Mind in the eyes and strange stories tasks (ToM measures): We used an abridged 12-trial version of the mind in the eyes task (Baron-Cohen et al., 1997, 2001). Participants were presented with pictures of someone's eyes and were asked to choose one word out of four choices that best described what the person was thinking or feeling (e.g., *amused*, *flustered*, *contemplative*, *encouraging*). The 12 trials were presented randomly, and the locations of the four answer choices were also randomized.

We also used an abridged version of the strange stories task (Happé, 1994) consisting of seven experimental and three control trials. Each of the experimental trials featured a short story describing a situation that involved pretend play, joking, white lies, figures of speech, irony, misunderstanding, or forgetting. Here is an example of such a story: "Katie and Emma are playing in the house. Emma picks up a banana from the fruit bowl and holds it up to her ear. She says to Katie, 'Look! This banana is a telephone!'" Participants were asked to explain why Emma said this, and the number of mental state (e.g., "She is pretending") and physical state (e.g., "A banana looks like a telephone") responses were tallied (coding was verified by a second rater, IRR = .91). Control stories did not involve mental states. A composite ToM measure was derived by calculating z-scores from the mean number of correct mind in the eyes trials and strange stories mental state justifications for each participant and then taking the sum of the two z-scores.

2.2. Results

Results of individual tasks are presented, followed by analyses targeting the relationships among EF, ToM, and pragmatic ability. Measures of reliability (split-half and Cronbach's alpha) for all tasks in Experiment 1 are presented below in Table 2. As can be seen, internal consistency was adequate, with all split-half correlations significant at $p < .001$ and Cronbach's alpha values at 0.704 and higher.

Table 2
Split-half reliability (Pearson correlations) and Cronbach's alpha measures for all tasks in Experiment 1

	Split-Half Reliability		Cronbach's Alpha
Task	*r*	*p*	α
Dual scalar implicature (ratings)	.944	<.001	0.942
Dual scalar implicature (memory accuracy)	.832	<.001	0.908
Auditory backward digit span	.869	<.001	0.906
Mind in the eyes	.497	<.001	0.705
Strange stories	.400	<.001	0.704
Simple scalar implicature	.646	<.001	0.728

2.2.1. *Dual scalar implicature task*

Results are presented in Fig. 2. A linear mixed-effects regression was performed on sentence rating data (excluding incorrect memory trials; 14.8% of the data on Load trials and 5.8% of the data on Control trials) using the *nlme* package (Pinheiro, Bates, DebRoy, & Sarkar, 2017) for the R Project for Statistical Computing v3.2.2 (R Core Team, 2015). Cognitive load (Control, Load), sentence type (Under-Informative, True [Some], True [All], False), and the interaction between the two were included in the model as fixed effects, with crossed random intercepts for Participants and Items. Sentence ratings differed significantly across sentence types, $\chi^2(3) = 6{,}953.639$, $p < .001$. Planned contrasts (presented in Table 3) indicated that Under-Informative ($M = 2.89$, $SD = 1.59$) sentences were rated higher than False ($M = 2.18$, $SD = 1.43$) sentences, $p < .001$, but lower than True (Some) ($M = 4.51$, $SD = 0.83$) sentences, $p < .001$. True (All) ($M = 4.34$, $SD = 1.05$) sentences were rated lower than True (Some) sentences, $p < .001$ (perhaps because participants thought that the quantifier *all* was superfluous). Importantly for present purposes, sentence ratings did not differ significantly between cognitive load conditions, $\chi^2(1) = 0.898$, $p = .343$, and the interaction between cognitive load and sentence type did not reach significance, $\chi^2(3) = 6.859$, $p = .077$.

2.2.2. *EF and ToM measures*

Results of all EF and ToM tasks are presented in Table 4.

2.2.3. *Simple scalar implicature task*

Performance on the simple scalar implicature task is shown in Fig. 3. As expected, participants were more likely to give "Bad" ratings to Under-Informative ($M = 3.32$, $SD = 1.96$) sentences than Informative ($M = 0.39$, $SD = 0.88$) ones, $t(177) = 17.74$, $p < .001$.[4]

We next related the number of "Bad" ratings of Under-Informative sentences on this task (henceforth, the *Pragmatic Score*) to performance on EF and ToM tasks. As a first step, we investigated the extent to which an individual's Pragmatic Score was correlated with their EF and ToM abilities. Pragmatic Score was significantly positively correlated with both composite EF scores, $\tau_b(176) = 0.142$, $p = .012$, and composite ToM scores,

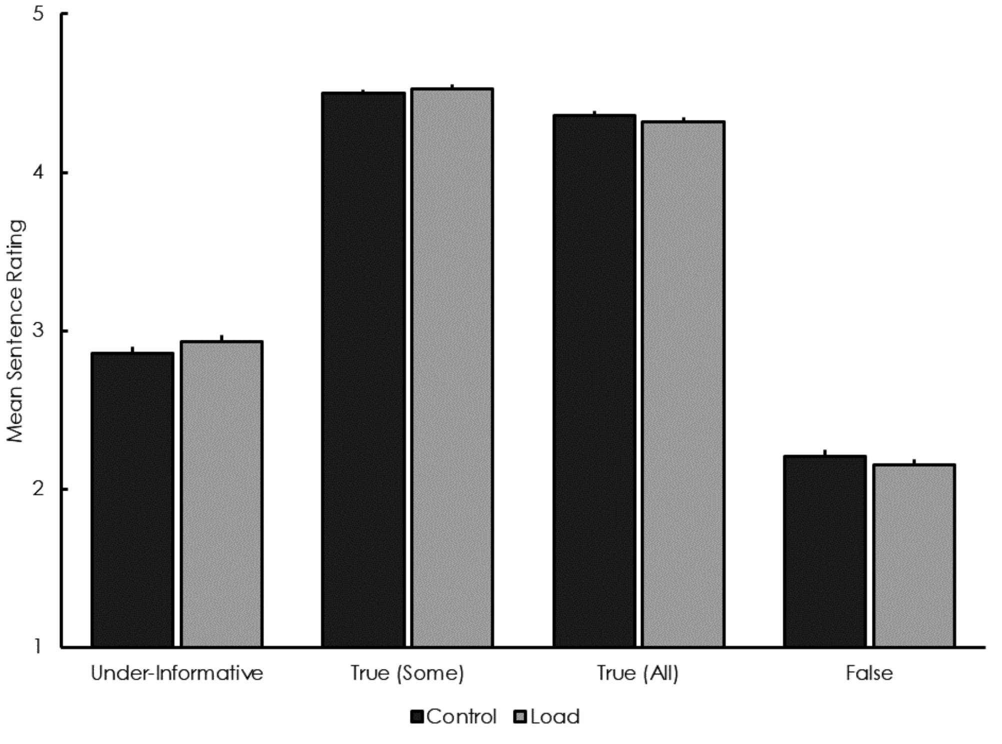


Fig. 2. Mean sentence ratings by sentence type and cognitive load condition in the dual scalar implicature task in Experiment 1. Error bars represent +1 *SEM*.

Table 3
Linear mixed-effects regression for the dual scalar implicature task results of Experiment 1, with cognitive load, sentence type, and their interaction included as fixed effects and crossed random intercepts for participant and item

Effect	β	*SE*	*t*	*p*
Intercept	3.475	0.039	88.269	<.001
Cognitive load (control vs. load)	−0.002	0.010	−0.220	.826
Sentence type (under-informative vs. false)	1.305	0.018	73.894	<.001
Sentence type (under-informative vs. true [some])	−0.275	0.020	−13.415	<.001
Sentence type (true [some] vs. true [all])	0.586	0.018	32.802	<.001

$\tau_b(176) = .203$, $p < .001$. These analyses align with prior work in demonstrating a role of each ability in scalar implicature, but they do not inform our understanding of the relative contribution of EF and ToM.

To tease apart the *unique* roles of EF and ToM in explaining the variation in judgments in the simple scalar implicature task, a multiple linear regression was conducted

Table 4
Scores on all EF and ToM measures in Experiment 1

Task	Mean	*SD*	Min	Max
Dual scalar implicature task				
Load trials memory accuracy	34.07	6.11	12	40
Backward digit span task				
Memory span	5.66	1.31	3	7
Mind in the eyes task				
No. correct	8.19	2.60	1	12
Strange stories task				
Mental state justifications	10.06	3.10	0	14
Physical state justifications	0.93	1.23	0	6
Control justifications	5.34	1.28	0	6

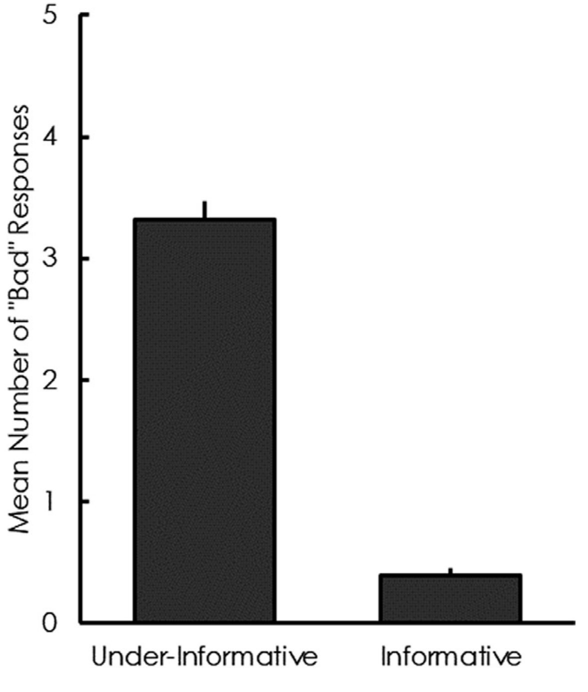


Fig. 3. Results of the simple scalar implicature task in Experiment 1. Error bars represent +1 *SEM*.

with the composite EF and ToM scores as independent variables and Pragmatic Score as the dependent variable. EF and ToM scores were significantly correlated in the present data set, $\tau_b(176) = 0.248$, $p < .001$, a common finding in the literature (e.g., Bull et al., 2008; Carlson et al., 2002) that further demonstrates the need to determine the unique impact of each factor on pragmatic ability, but the moderate correlation did not raise any issues of multicollinearity (all VIF values < 1.5, all tolerance values < 1) so the regression was performed as planned. The model accounted for a significant amount of

Table 5
Multiple linear regression predicting pragmatic score (the number of "bad" ratings of under-informative sentences in the simple scalar implicature task) in Experiment 1 from EF and ToM scores

Effect	β	*SE*	*t*	*p*
Intercept	3.320	0.140	23.659	<.001
EF	0.128	0.092	1.386	.167
ToM	0.329	0.096	3.429	<.001

variance, $F(2, 175) = 9.878$, $p < .001$, $R^2 = 0.09$. As can be seen in Table 5, EF was not significantly associated with Pragmatic Score, $p = .167$. ToM was significantly positively associated with Pragmatic Score, $p < .001$. In other words, participants who performed better on ToM tasks penalized Under-Informative sentences more on the simple scalar implicature task, and this relationship held even when controlling for EF. In contrast, EF had no unique impact on the responses on the simple scalar implicature task.[5,6]

For our last analysis, we subtracted the number of times Informative sentences were judged as "Bad" by an individual participant from the number of times Under-Informative sentences were judged as "Bad." Calculating this pragmatic difference score (PDS) allowed us to account for individual responding preferences (i.e., the general likelihood that a participant would judge a sentence as "Bad"), and thus gave a more sensitive picture of pragmatic sensitivity. The maximum PDS of 5 represents a completely pragmatic participant who would always reject Under-Informative statements and accept Informative statements. A score of 0 represents a completely logical participant who would always accept both Under-Informative and Informative statements. Finally, a highly unlikely negative score would be indicative of a participant who would judge Informative statements to be "Bad" more often than Under-Informative statements.

We repeated the previous analyses using PDS as the dependent variable instead of Pragmatic Score (see Table 6). PDS was significantly positively correlated with both composite EF scores, $\tau_b(176) = 0.179$, $p = .001$, and composite ToM scores, $\tau_b(176) = 0.257$, $p < .001$. To investigate the *unique* roles of EF and ToM in explaining the variation in judgments in the simple scalar implicature task, a multiple linear regression was conducted with the composite EF and ToM scores as independent variables and PDS as the dependent variable. The model accounted for a significant amount of variance, $F(2, 175) = 18.250$, $p < .001$, $R^2 = 0.16$. EF was not significantly associated with

Table 6
Multiple linear regression predicting pragmatic difference score (the number of "bad" ratings of under-informative sentences minus the number of "bad" ratings of informative sentences in the simple scalar implicature task) in Experiment 1 from EF and ToM scores

Effect	β	*SE*	*t*	*p*
Intercept	2.927	0.151	19.387	<.001
EF	0.173	0.099	1.741	.083
ToM	0.491	0.103	4.754	<.001

PDS, $p = .083$, but ToM was significantly positively associated with PDS, $p < .001$. Thus, our finding that participants who performed better on ToM—but not EF—tasks behaved more pragmatically on the simple scalar implicature task was demonstrated again even when controlling for response preferences by using PDS.[7]

2.3. Discussion

The main goal of Experiment 1 was to test the contributions of EF and ToM in explaining variability in scalar implicature judgments. As a first step toward this goal, in a dual scalar implicature task we attempted to replicate the previous finding that individuals make more literal judgments under greater cognitive load (De Neys & Schaeken, 2007) due to the recruitment of EF in scalar implicature computation. We did not observe such a pattern of results in the present study: Judgments of Under-Informative sentences did not differ significantly between high memory load and control conditions.

Our next step involved relating individual scores on EF and ToM tasks to judgments in a simple scalar implicature task. This step went beyond much prior work on pragmatics—including the De Neys and Schaeken (2007) study—that typically investigated only one ability (EF, in this case) at a time. Even though both EF and ToM were significantly correlated with scalar judgments, only ToM had a unique association with such judgments: Participants who performed better on the ToM tasks were less literal in their interpretation of Under-Informative utterances, more often making judgments consistent with the calculation of a (potential) scalar implicature. This association is predicted by a wide class of pragmatic accounts inspired by Grice (1975), which require a listener to reason about the intentions of the speaker in order to successfully compute the implicature.

Why did our dual task results differ from those in De Neys and Schaeken's study (and other prior work such as Marty & Chemla, 2013)? Our study tested a larger number of participants (179 compared to 56) and included more trials (20 per condition for a total of 80 as compared to 10 per condition for a total of 20 in the original study), so it presented a strong test of the hypothesis that motivated the earlier work. There are some methodological differences between the studies but these cannot readily explain the divergence in results. For instance, our paradigm, unlike the earlier study, allowed participants to take as much time as they needed when inspecting the dot patterns, whereas, in previous studies, the dot patterns were presented for only 850 ms (participants could take as long as they needed to rate the sentence and recreate the dot pattern both in our work and in prior studies). Nevertheless, there is no evidence that this led to reduced load demands: In fact, our task yielded more memory errors on Load trials (14.8% compared to 6% in De Neys and Schaeken's study). Furthermore, as Table 4 shows, there was a range in working memory performance that could support Load differences (cf. Dieussaert et al., 2011).

The biggest difference from earlier work is that our paradigm used a 5-point rating scale instead of a binary truth-value judgment task. One might imagine that, if participants overwhelmingly chose middle values on the 5-point scale, there would be no basis for expecting load effects (on such middle choices, see Katsos & Bishop, 2011; Schaeken et al., 2018). However, a closer inspection of our findings shows that participants gave a continuum of more versus less pragmatic responses, as opposed to clustering around middle scores. For instance, within each of the Control and Load conditions, only 48 out of a total of 178 participants gave an average middle rating (between 2.5 and 3.5) to Under-Informative sentences. Given this continuum, it is not clear why the scale by itself should have removed the EF effect: If EF uniquely bears on implicature computation, it should still be the case that more pragmatic ratings on a 5-point scale should correspond with higher EF capabilities, and vice versa. Notice also that the simple scalar implicature task involved a binary scale but also offered no evidence for a unique EF contribution (even though EF was correlated with pragmatic responding in that task). Even considering these methodological differences, then, our data challenge the conclusion that individuals are more logical in interpreting linguistic meaning under greater cognitive load.

More importantly, our data show that, even when a relationship between EF and scalar implicature is found in individual responders such as in our correlation analyses in the simple scalar implicature task, this relationship disappears when ToM is accounted for. The same analysis applied to the results from the dual scalar implicature task confirms this conclusion. Previous work, including the studies conducted by De Neys and Schaeken (2007) and Marty and Chemla (2013), did not measure ToM and therefore missed a critical contributor to pragmatic inference. It is therefore possible that previously observed contributions of EF could be reinterpreted in the context of the role of ToM.[8] We revisit this issue in Experiment 2.

3. Experiment 2: Computing metaphors, indirect requests, and scalar implicatures

In Experiment 1, we found that ToM—but not EF—was important in making pragmatic judgments. To test whether this association is specific to scalar implicature or generalizes to other pragmatic domains, in Experiment 2 we investigated the contributions of EF and ToM to metaphor and indirect request, as well as scalar implicature (to replicate the findings of Experiment 1).

Metaphors ("I'm a night owl") require the listener to infer that the speaker did not intend to convey the literal meaning of the utterance (as that meaning violates the maxim of quality—it is untrue) but rather some kind of similarity (in the above example, staying up late; Grice, 1975). As common metaphors become familiar/conventionalized, the pragmatic meaning may become automatically available (Glucksberg, 2003). Indirect requests ("It's hot in here") require the listener to infer that the speaker did not intend to convey simply the literal meaning of the utterance (since this would violate the maxim of quantity—this information is obvious and known to both speaker and listener; Grice, 1975); the purpose of the utterance is to indirectly ask the hearer to perform an action that he/she is willing and able to do (e.g., open the window). There is evidence that both EF and ToM are recruited in both metaphors and indirect requests. Behavioral work measuring individual differences in working memory has demonstrated an association between metaphor comprehension and EF (Chiappe & Chiappe, 2007; Mashal, 2013), and ERP research suggests that processing indirect requests increases higher memory retrieval demands as compared to processing non-request utterances (Coulson & Lovett, 2010). Furthermore, metaphor comprehension has been associated with ToM skills (Happé, 1993; but see Norbury, 2005), and engagement of ToM regions of the brain has been reported for both metaphor (Prat et al., 2012) and indirect request comprehension (Van Ackeren et al., 2012). However, the specific contribution of EF and ToM to pragmatic variation across individuals has not been tested yet.

To assess metaphor comprehension, we borrowed a task from Jankowiak, Rataj, and Naskręcki (2017). That study investigated the neural correlates of processing novel metaphors ("to harvest courage"), conventional metaphors ("to gather courage"), literal phrases ("to experience courage"), and anomalous phrases ("to move courage") in Polish-English bilinguals' first and second languages. (For ease of presentation and because our participants were all monolingual, we focus here on the first-language results.) The authors reasoned that novel metaphors would require more cognitive resources to be processed than conventional metaphors because understanding a novel metaphor requires a comparison between the literal and the intended meaning (Bowdle & Gentner, 2005), whereas understanding more familiar, conventional metaphors does not (Glucksberg, 2003). In line with this prediction, novel metaphors elicited greater N400 components than conventional metaphors, literal phrases, and anomalous phrases, indicative of the increased semantic processing necessary to comprehend them. Interestingly, novel metaphors were also judged as less meaningful than conventional metaphors, and conventional metaphors as less meaningful than literal phrases (anomalous phrases were the least meaningful items). Here we adopted the behavioral paradigm of Jankowiak et al. (2017) and asked participants to judge the meaningfulness of metaphorical and literal phrases, expecting to replicate the finding that novel metaphors are judged as less meaningful than literal phrases overall. Of interest was whether the gap between literal and metaphorical phrases would be narrower for participants with higher ToM (indicating successful comprehension of the pragmatic meaning of metaphors).

To test indirect request comprehension, we used a paradigm originally developed by Van Ackeren et al. (2012) to investigate the activation of ToM regions of the brain during the processing of indirect requests. In this task, participants were presented with picture–sentence pairs that either suggested that the speaker was making a request (e.g., a picture of a closed window paired with the sentence "It is very hot here") or not (e.g., a picture of a desert paired with the sentence "It is very nice here"). Compared to control trials, request trials elicited greater activation of cortical regions associated with ToM. Furthermore, behavioral judgments indicated that participants were more likely to feel that the speaker was making a request in the critical indirect request condition as compared to control trials. Here we administered a behavioral adaptation of the task using the same stimuli and asked participants to rate how strongly they felt that the speaker was making a request. We expected such ratings to be higher for request trials as compared to control trials. Of interest was whether better ToM (after controlling for EF) would enhance this pattern.

3.1. Method

3.1.1. Participants

Two hundred monolingual English speakers aged 19–68 (*M* = 29.38, *SD* = 5.66) living in the United States, 91 of whom were female, were recruited from Amazon's Mechanical Turk to participate in Experiment 2. Participants were compensated $1.50 for the 15-min study. Data from 26 individuals who reported to be bilingual and/or diagnosed with ASD were excluded, leaving 174 participants for analysis.

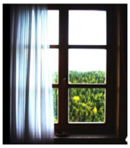
3.1.2. Materials and procedure

Participants completed five tasks in the order presented below: three pragmatic tasks (metaphor, indirect request, scalar implicature), an auditory backward digit span task (as a measure of EF), and the mind in the eyes and strange stories tasks (as measures of ToM). We used a single measure of EF here because the second measure in Experiment 1 would have been impractical to include and did not change the pattern of findings (recall that we analyzed the results for Experiment 1 using this measure of EF alone and results remained the same). Stimuli for the metaphor and indirect request tasks can be found in the Supplementary Material.

3.1.2.1. Metaphor task: Participants were presented with 20 verb phrases individually and were asked to rate on a 5-point scale how meaningful each one was. Ten phrases were novel metaphors ("to harvest courage") and 10 literal phrases ("to feel anger"). The materials (borrowed from Jankowiak et al., 2017) had been extensively normed on word length, concreteness, metaphoricity, etc. in the earlier study. Unlike Jankowiak et al. (2017), we did not include conventional metaphors or anomalous phrases as we were interested in the clearest contrast between nonliteral/metaphorical and literal phrases.

3.1.2.2. Indirect request task: Participants were presented with 36 picture–sentence pairs borrowed from Van Ackeren et al. (2012), evenly divided over four types of trials (see Table 7). On the critical indirect request trials, the sentences accompanying the pictures (e.g., the sentence "It is very hot here" paired with a picture of a closed window) could be used to imply that the speaker wished the listener to do something (e.g., open the window)—that is, they could be taken as an indirect request. The three control conditions (picture control, utterance control, and picture-utterance control) were designed in such a way as to discourage a request interpretation (see Table 6). The picture control trials combined images from the indirect request trials with an utterance that could not readily be interpreted as a request. The utterance control trials combined the request statements from the indirect request trials with a picture that no longer provided supporting context for a request interpretation. The picture-utterance control trials combined the images from the utterance control trials with the statements from the picture control trials, the combination of which again discouraged a request interpretation. On each trial, the sentence appeared beneath the picture and participants were asked, "How much do you feel that

Table 7
Examples of stimuli used in the indirect request task in Experiment 2 (reprinted with permission from Van Ackeren et al., 2012)

Picture	Sentence	Trial Type
	It is very hot here It is very nice here	Indirect request Picture control
	It is very hot here It is very nice here	Utterance control Picture-utterance control

the speaker wants something from you?" They were given a 5-point scale to input their responses and were instructed that a higher rating meant that they felt strongly that the speaker wanted something from them. Trials were presented in a random order for each participant, and stimuli were fully rotated such that only one version of a given item was presented to each participant. Higher ratings on indirect request trials would indicate pragmatic (request) interpretations.

3.1.2.3. Scalar implicature task: This task was identical to the simple scalar implicature task administered in Experiment 1.

3.1.2.4. Auditory backward digit span task: This EF measure was identical to the task in Experiment 1, and the *z*-score of participants' mean performance was used as the sole measure of EF in Experiment 2.

3.1.2.5. Mind in the eyes and strange stories tasks: These ToM measures were the same as in Experiment 1. Coding for the strange stories task was verified by a second rater ($IRR = .89$). The composite ToM score was calculated in the same way as in Experiment 1.

3.2. Results

Results for the metaphor, indirect request, and scalar implicature task are given in Fig. 4. Overall results from EF/ToM tasks are given in Table 8. In what follows, we analyze results separately for each pragmatic task and relate them to EF and ToM measures.

Measures of reliability (split-half and Cronbach's alpha) for all tasks in Experiment 2 are presented in Table 9. As in Experiment 1, internal consistency was adequate, with split-half correlations all significant at $p < .001$ and Cronbach's alpha values of 0.637 and higher.

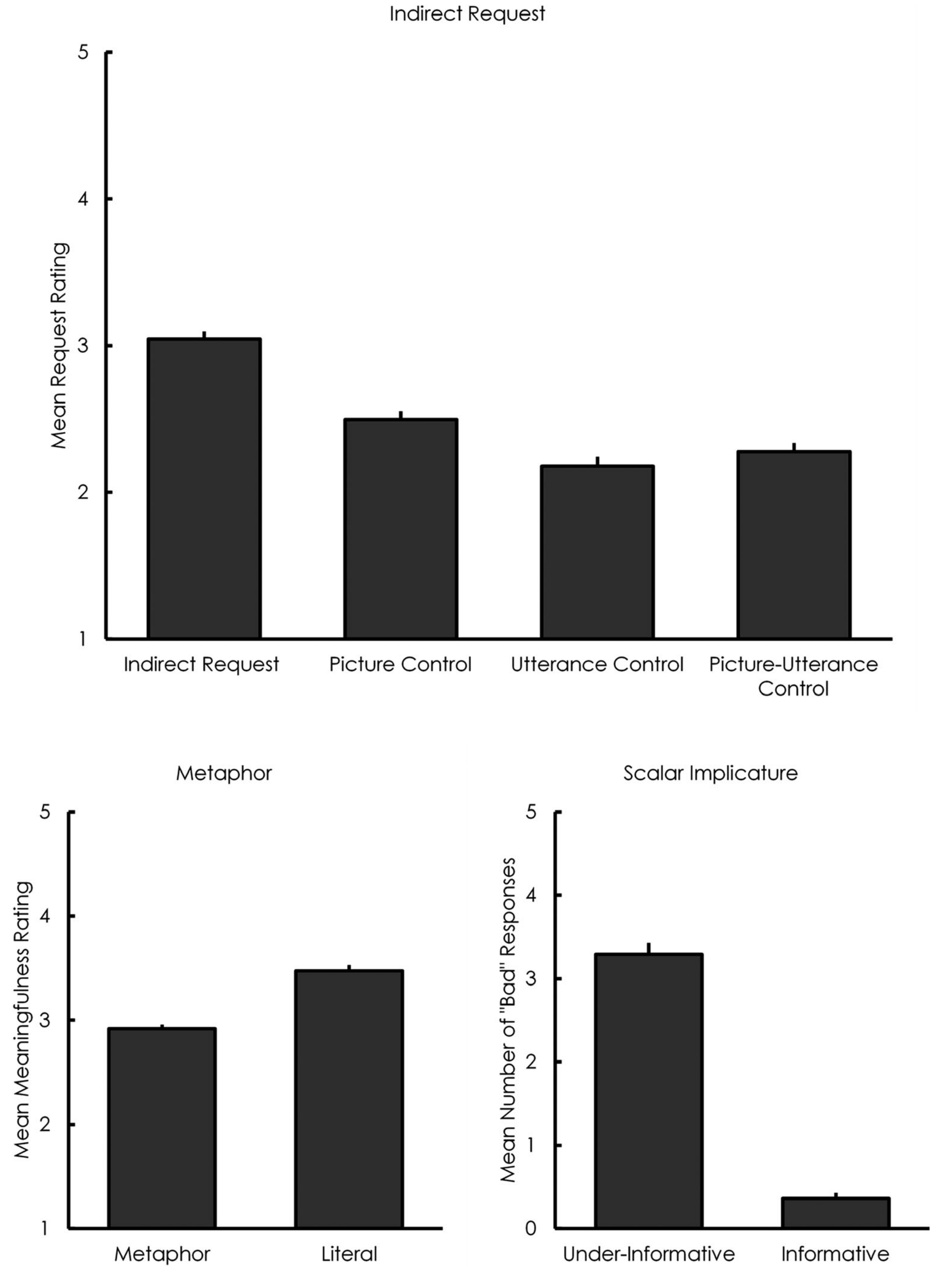


Fig. 4. Results of the metaphor, indirect request, and scalar implicature tasks in Experiment 2. Error bars represent +1 *SEM*.

Table 8
Scores on all EF and ToM tasks in Experiment 2

Task	Mean	*SD*	Min	Max
Backward digit span task				
Memory span	5.55	1.53	3	7
Mind in the eyes task				
No. correct	8.77	2.47	0	12
Strange stories task				
Mental state justifications	10.57	2.80	0	14
Physical state justifications	1.17	1.54	0	6
Control justifications	5.45	1.17	0	6

3.2.1. Metaphor task

As shown in Fig. 4, participants judged literal control phrases ($M = 3.48$, $SD = 0.69$) as being more meaningful than the critical metaphorical statements ($M = 2.92$, $SD = 0.50$), $t(173) = 8.846$, $p < .001$, as predicted. As before, we first investigated the extent to which an individual's responses were correlated with their EF and ToM abilities. Meaningfulness ratings on critical trials of the metaphor task were not significantly correlated with EF scores, $\tau_b(173) = -0.041$, $p = .483$, but somewhat counterintuitively were significantly *negatively* correlated with ToM scores, $\tau_b(173) = -0.176$, $p < .001$.

We then conducted a regression analysis predicting pragmatic ability in the metaphor task from EF and ToM to determine the unique contribution of each. As in Experiment 1, to make results easier to compare across tasks and to account for general response preferences, the regression analysis was based on a difference score that sought to measure the level of pragmatic responding tendencies given responses to non-pragmatic trials. Specifically, for each participant, we used the mean metaphorical phrase ratings minus the mean control (literal) phrase ratings to construct a PDS: More negative values represent more logical responding (i.e., treating literal phrases as more meaningful than metaphorical ones), whereas values closer to zero indicate that the metaphors were judged more closely to literal statements in terms of meaningfulness, presumably because their enriched, non-literal meanings were available to the comprehender. A PDS score of 0 represented an individual who took metaphorical statements to make as much sense as literal ones. A

Table 9
Split-half reliability (Pearson correlations) and Cronbach's alpha measures for all tasks in Experiment 2

	Split-Half Reliability		Cronbach's Alpha
Task	*r*	*p*	α
Metaphor	.563	<.001	0.646
Indirect request	.840	<.001	0.913
Simple scalar implicature	.618	<.001	0.648
Auditory backward digit span	.910	<.001	0.938
Mind in the eyes	.562	<.001	0.708
Strange stories	.422	<.001	0.637

multiple linear regression with EF and composite ToM scores as predictor variables and PDS as the outcome variable did not reach significance, $F(2, 171) = 3.018$, $p = .052$.

3.2.2. Indirect request task

Results are shown in Fig. 4. A linear mixed-effects regression performed on responses in the indirect request task with trial type (indirect request, picture control, utterance control, picture-utterance control) included in the model as fixed effects and crossed random intercepts for participants and items indicated that the extent to which participant felt that a request was being made differed significantly across trial types, $\chi^2(3) = 254.061$, $p < .001$. Post-hoc tests indicated that indirect request ($M = 3.04$, $SD = 0.69$) trials were rated higher than picture control ($M = 2.18$, $SD = 0.83$), utterance control ($M = 2.50$, $SD = 0.76$), and picture-utterance control ($M = 2.28$, $SD = 0.77$) trials, all ps < .001. Picture control trials were rated lower than utterance control and picture-utterance control trials, both ps < .001. Picture-utterance control and utterance control trials did not differ significantly from one another, $p = .060$. Request ratings on the critical trials of the indirect request task were not significantly correlated with EF scores, $\tau_b(173) = -0.021$, $p = .722$, or ToM scores, $\tau_b(173) = -0.056$, $p = .284$.

As with metaphor, we created a PDS that corresponded to the mean request rating in the critical request trials minus the mean request rating across all control trials for each participant. The maximum PDS of 4 represented perfect identification of indirect requests and non-requests, whereas smaller values indicated an inability to detect indirect requests. A multiple linear regression predicting PDS from EF and composite ToM scores was significant, $F(2, 171) = 17.700$, $p < .001$, $R^2 = 0.17$ (Table 10). The association between EF and PDS was not significant, $p = .518$. ToM was positively associated with PDS, $p < .001$, such that participants who performed better on the ToM tasks were also more accurate at identifying an indirect request.

3.2.3. Scalar implicature task

As shown in Fig. 4, participants were more likely to give "Bad" ratings to Under-Informative ($M = 3.29$, $SD = 1.81$) sentences than Informative ($M = 0.36$, $SD = 0.89$) ones, $t(173) = 18.149$, $p < .001$. The number of "Bad" ratings to Under-Informative sentences was not significantly correlated with EF scores, $\tau_b(173) = 0.108$, $p = .084$, but it was significantly positively correlated with ToM scores, $\tau_b(173) = 0.199$, $p < .001$.

Table 10
Multiple linear regression analyses predicting pragmatic difference scores (PDSs) on the indirect request and scalar implicature tasks in Experiment 2 from EF and ToM scores

	Indirect Request		Scalar Implicature	
Effect	β (*SE*)	*p*	β (*SE*)	*p*
Intercept	0.726 (0.053)	<.001	2.925 (0.149)	<.001
EF	0.037 (0.056)	.518	0.086 (0.160)	.591
ToM	0.177 (0.023)	<.001	0.473 (0.094)	<.001

We computed a PDS for this task corresponding to the number of “Bad” ratings for Under-Informative sentences minus the number of such ratings for Informative sentences for each participant. As in Experiment 1, the maximum PDS of 5 represented a perfectly pragmatic responder who always highly penalized Under-Informative sentences but never Informative sentences. A multiple linear regression was then performed with EF and composite ToM scores as predictor variables and PDS as the outcome variable to determine the unique influence of each predictor on scalar implicature judgments. Overall, the analysis predicted a significant amount of variance on the scalar implicature task, $F(2, 171) = 15.640$, $p < .001$, $R^2 = 0.14$ (Table 10). The association between EF and PDS was not significant, $p = .591$, but ToM was positively associated with PDS, $p < .001$.

3.3. Discussion

In Experiment 2, our primary aim was to investigate the relationship between performance on a broad array of pragmatic tasks (scalar implicature, metaphor, and indirect request judgments) and measurements of EF and ToM within individuals. ToM—but not EF—was significantly positively associated with indirect request and scalar implicature judgments (as captured by PDSs), replicating and extending Experiment 1. For metaphor, the main analysis did not yield significant results. In simple correlation analyses, EF was not predictive of participants' judgments of metaphorical statements, but ToM was negatively associated with pragmatic responding on the metaphor task.

Our findings from the metaphor task appear particularly counterintuitive given the link observed here and elsewhere between better ToM skills and enhanced pragmatic abilities (Chiappe & Chiappe, 2007; Happé, 1993; Mashal, 2013; even though see Norbury, 2005). Notice that, in both Jankowiak et al. (2017) and our own data, metaphorical phrases overall were judged as less meaningful than literal ones. This is unsurprising since such phrases had both literal (false) and metaphorical (true) meanings, unlike literal phrases that only had one true meaning (recall also that under-informative statements were typically rated higher than completely false statements but lower than completely true statements, as listeners take into account both the false semantic and true pragmatic meanings—see Experiment 1, dual scalar implicature task; cf. Fairchild & Papafragou, 2018; Katsos & Bishop, 2011). What *is* surprising and requires an explanation is the fact that in the present study the literal-metaphorical ratings gap did not change (and in fact, according to the simple correlation analysis, became *wider*) with increased ToM sophistication.

A possible explanation for this outcome is that the task invited a search for criteria for meaningfulness, and that participants with greater metacognitive awareness were more inclined to understand the task as requiring a distinction between canonically meaningful (literal/semantic) and non-canonically meaningful (figurative/pragmatic) items. As the negative correlation shows, participants with better ToM gave lower ratings to phrases that required extra processing to become meaningful, as compared to literal phrases that were already meaningful at the semantic level. In other words, better ToM capacities were associated with increased awareness that the meaningfulness of the metaphorical phrases is figurative, instead of literal. Notice, however, that this association between ToM and metaphor ratings disappeared once we controlled for EF.

Two further considerations support this line of reasoning. First, the metaphor task involved judging the meaningfulness of verb phrases in the infinitival form ("to do X"), a fact that may have made participants more likely to focus on the phrases as linguistic tokens and apply their mindreading skills to understand the intent of the task. This is in contrast to our indirect request task, for example, where participants were presented with whole utterances embedded in a context and explicitly asked to reason about what the speaker wanted/intended (cf. also other studies that tested for metaphor comprehension; Chiappe & Chiappe, 2007; Gentner, 1988; Mashal, 2013). Second, even literal phrases elicited relatively low scores in our task compared to Jankowiak et al. (2017) and other work (Bambini, Ghio, Moro, & Schumacher, 2013). This may be due to the fact that, for simplicity, our design omitted conventional metaphors and anomalous phrases that were included in the original study.

While we leave open these possibilities concerning the metaphor task, the results of Experiment 2 overall highlight the role of ToM in pragmatic processing. The fact that we observed very similar results for scalar implicature and indirect request comprehension suggests that underlying ToM mechanisms are at play across distinct pragmatic phenomena. Even in the case of metaphor, pragmatic judgments within individuals seem to depend on ToM abilities, even though this effect is somewhat different from what was anticipated and disappears once we control for contributions from EF.

4. General discussion

Adults are often used as the benchmark for pragmatic computation in both developmental studies (Katsos & Bishop, 2011; Noveck, 2001; Papafragou & Musolino, 2003) and comparisons with atypical populations (Chevallier et al., 2010; Hochstein et al., 2017; Noveck et al., 2007; Pijnacker et al., 2009), yet there is tremendous variation in (neurotypical) adults' pragmatic judgments. Here we explored the roots of this variation focusing on the well-studied case of scalar implicature. Many psycholinguistic investigations of this phenomenon have noted individual differences when examining adults' pragmatic performance (Bott & Noveck, 2004; Degen & Tanenhaus, 2015; Fairchild & Papafragou, 2018; Heyman & Schaeken, 2015; Noveck & Posada, 2003; Politzer-Ahles et al., 2013); nevertheless, the underlying cause for such individual differences is not clear. Furthermore, it is unknown whether an individual's ability to compute scalar implicatures is linked to the ability to compute other kinds of pragmatic implicatures. Our goal in the present paper was to address these gaps in the literature.

Drawing on the theoretical assumption that deriving a scalar implicature can be cognitively costly (e.g., Geurts, 2010; Grice, 1975; Sauerland, 2004; Sperber & Wilson, 1986), and prior empirical findings (Antoniou et al., 2016; De Neys & Schaeken, 2007, among others), we tested the hypothesis that EF—specifically working memory—might be associated with scalar implicature calculations. Similarly, drawing on theoretical assumptions about the process of scalar implicature computation (Grice, 1975; Sauerland, 2004; Sperber & Wilson, 1986) and previous experimentation (Noveck et al., 2007, among others), we tested the hypothesis that ToM might be associated with scalar pragmatic judgments. In Experiment 1, we presented the first investigation to incorporate measures of both EF and ToM in an attempt to tease apart the unique impact of the two factors on scalar implicature calculations within an individual. In Experiment 2, we extended our investigation to metaphor and indirect request in an effort to understand whether the mechanisms underlying individual differences in scalar implicature computation extend to other types of pragmatic phenomena.

Across the two experiments, we found that the better a participant performed on tasks measuring ToM, the more likely they were to respond to the pragmatic meaning (as opposed to the literal meaning) of an under-informative utterance such as "Some dogs are mammals." Importantly, this association represented the unique influence of ToM on scalar implicature judgments, controlling for EF. In addition to the strong and replicable association between ToM and scalar implicature judgments, we found that—after controlling for EF—participants with stronger ToM were better able to identify when a speaker was making an indirect request. Interestingly, we saw no such effect for metaphor. In fact, in a simple correlation analysis, we found a moderate association in the opposite direction for metaphor, such that participants with better ToM displayed higher meaningfulness ratings for *literal* compared to metaphorical sentences (presumably because they interpreted the task as targeting literal meaningfulness). In terms of measures of EF (working memory), we failed to find a unique relationship between EF and pragmatic judgments of any type (including scalar implicature, metaphor, and indirect request), suggesting that—at least for adults—one's cognitive resources are not uniquely important for making pragmatic judgments.

4.1. Executive function and pragmatic computations

Our experiments reveal no unique contribution of EF to scalar implicature computation once ToM has been factored in (even though there were significant positive correlations between EF and performance on scalar implicature; see Experiment 1). We therefore conclude that, to the extent that EF bears on pragmatic implicature computation, it does so via ToM reasoning.

There are two specific, not mutually exclusive possibilities about the role of EF in ToM reasoning during pragmatic interpretation. One possibility is that EF is recruited during ToM computations because these computations are inherently effortful (Apperly, Back, Samson, & France, 2008; Apperly, Samson, Chiavarino, & Humphreys, 2004; Keysar, Barr, Balin, & Brauner, 2000; Lin, Keysar, & Epley, 2010). Some evidence supporting this possibility comes from referential communication paradigms in which adults are slower to select an appropriate referent when they need to incorporate another person's perspective (Epley, Morewedge, & Keysar, 2004; Keysar et al., 2000). Additionally, dual-task studies demonstrate that adults have difficulty selecting the appropriate visual perspective in a perspective-taking task (but not performing simple level-1 visual perspective calculations; Qureshi, Apperly, & Samson, 2010[9]) and completing the very same mind in the eyes and strange stories tasks we administered here (Bull et al., 2008) when EF demands are high. According to this possibility, previously observed associations between EF and scalar implicature (Antoniou et al., 2016; De Neys & Schaeken, 2007) may have been actually indicative of the link between EF and ToM. For instance, individuals under heavy cognitive load in previous work may have had fewer cognitive resources available to engage in ToM and reason about the intended meaning of the under-informative sentences, and thus responded more often to the literal meanings.

Alternatively, EF might be required to hold the products of ToM reasoning in memory and combine them with other information such as linguistic context during the computation of pragmatic meaning (cf. Antoniou et al., 2016). In the specific case of scalar implicature ("Some dogs are mammals"), participants needed to combine their reasoning about what a cooperative, knowledgeable interlocutor would say with their knowledge of the world (e.g., that dogs are mammals, and that there are no exceptions). Thus, participants who failed to derive a scalar implicature (or otherwise limited their responses to logical content) may have done so because they did not have the EF resources to engage in ToM and/or connect that ToM reasoning with their own world knowledge. This explanation applies both to our own work and to prior work (e.g., De Neys & Schaeken, 2007): When listeners' EF was taxed from heavy working memory load, listeners may have had difficulties not simply in engaging in ToM, but also in comparing the result of this reasoning with their own world knowledge and the activation of the stronger alternative utterance (e.g., "All dogs are mammals").

Given our findings, the specific subcomponent(s) of EF involved in pragmatic inference should be the topic of further investigation. There is some disagreement as to whether working memory is synonymous with EF or is simply one of its subcomponents, along with inhibition and task-switching (Miyake et al., 2000, who also points out that there are additional EF components; see also Cooper, 2010, and others). Some of this disagreement may be attributed to the wide variety of tasks used to measure EF, which often suffer from issues of reliability and validity (Duckworth & Kern, 2011). In both of our experiments, we measured EF through the backward digit span task used either alone (Experiment 2) or in combination with a visual working memory measure (Experiment 1). Future research could include additional tasks targeting additional EF subcomponents.

4.2. Theory of mind and pragmatic computations

A key aspect of our findings is that scalar implicature and indirect request calculations in neurotypical adults (drawn from the general population of participants in online studies) are related to these adults' ToM skills, even when controlling for EF. This novel finding is in line with previous research reporting impaired pragmatic reasoning in patients with cortical lesions to ToM areas (Champagne-Lavau & Joanette, 2009; Spotorno, Koun, Prado, Van Der Henst, & Noveck, 2012; cf. also Wampers, Schrauwen, De Hert, Gielen, & Schaeken, 2018) and increased neural activation of ToM areas in healthy adults when processing indirect requests (Van Ackeren et al., 2012). Taken together, this evidence supports the claim that understanding the pragmatic meaning of an utterance requires actively thinking about what the speaker intended to say—a claim with a long history within the linguistic and philosophical literature on pragmatics (Geurts, 2010; Grice, 1975; Sauerland, 2004; Sperber & Wilson, 1986). (On the somewhat different pattern of results on metaphor, see Section 3.3.)

One might be surprised by the role of ToM in our data since the present tasks included no actual conversations and no specific speakers whose mental state participants could engage with. Indeed, several researchers have raised the possibility that some pragmatic inferences might not require ToM (Hochstein et al., 2017; for discussion, see also Heyman & Schaeken, 2015; Pijnacker et al., 2009; Wampers et al., 2018). Here we adopt the position that utterances invite assumptions about the conversational context in which they could reasonably have been produced, even when no observable speaker is present (Grice, 1975; Sperber & Wilson, 1986). The level of granularity in the hearer's model of the speaker can vary considerably. For the present simple scalar implicature task, assumptions about the speaker involve generic information about what is known within a community, including scalar alternatives that could have been used but were not. For the indirect request task, assumptions about the speaker are additionally constrained by visual cues about where the conversation took place (e.g., a room with a closed window). For the metaphor task, for reasons associated with specific task demands, assumptions about the speaker are more diffuse and do not uniquely enter into the interpretation of how meaningful novel metaphors are. If this line of reasoning is correct, the role of ToM should generalize to pragmatic tasks that include richer speaker information (e.g., reasons for producing under-informative utterances: Bonnefon, Feeney, & Villejoubert, 2009; Fairchild, Mathis, & Papafragou, 2020).

Our findings confirm the presence of substantial variation across individuals in reasoning about others' minds. It is important to consider, however, that, much like EF, ToM is a broad term for a collection of abilities relating to thinking about other people, including perspective-taking, false belief understanding, emotion recognition, and so forth. The tasks we administered to measure ToM both required a high degree of linguistic ability—vocabulary knowledge for the mind in the eyes task and the ability to synthesize multiple pieces of information embedded in a story and generate a response based on this information for the strange stories task. We chose these tasks because they are more complex and particularly well suited for use with adults (Baron-Cohen et al., 2001; Happé, 1994; Jolliffe & Baron-Cohen, 1999). It would be interesting to extend our approach to include established ToM tasks with reduced linguistic demands (see, e.g., Samson, Apperly, Braithwaite, & Andrews, 2010).

Our results leave it open whether ToM is recruited online during the early stages of pragmatic inference, or at a later stage of reasoning before providing a judgment. Online measures of scalar implicature computation (e.g., Breheny et al., 2006; Huang & Snedeker, 2009; Nieuwland, Ditman, & Kuperberg, 2010), EF (e.g., Eriksen & Eriksen, 1974), and ToM (e.g., Dumontheil, Apperly, & Blakemore, 2010) can be integrated into future work to help tease these two possibilities apart (cf. Breheny et al., 2013, for evidence that speaker knowledge is consulted early during the processing of scalar terms).

4.3. *Pragmatic performance within and across tasks*

The central motivation for the present research was the observation that adults vary in how consistently they make pragmatic inferences. This variation has been dealt with in the past by splitting participants into logical and pragmatic responders (e.g., Bott & Noveck, 2004). Here, we chose to treat pragmatic responding as a continuum rather than a dichotomy given inconsistencies within individuals observed in previous research (Degen & Tanenhaus, 2015; Fairchild & Papafragou, 2018), as well as in the present data, where judgments spanned the entire scale of possible responses. Regardless of how variation in pragmatic inference is treated, it is important to consider the issue of stability of responding—whether "pragmatic" responders generally behave pragmatically, for example, across contexts.

Our findings point to differences in adults' pragmatic response patterns based on (presumably stable) participant-internal cognitive characteristics (related specifically to the individual's ToM). Nevertheless, as the metaphor data show, this picture is far from simple. Other data in the present study suggest that participant-external factors—especially, the nature of the task—are likely to shift participant-internal biases. For instance, judgments on the two scalar implicature tasks were only moderately correlated in Experiment 1 (see note 4). This result is reminiscent of studies highlighting the influence of context and task demands on scalar implicature computation (Bergen & Grodner, 2012; Bonnefon et al., 2009; Breheny et al., 2006; Huang & Snedeker, 2018) and points to the need to integrate participant-driven and task-driven factors in understanding variation in adults' pragmatic responding within and across tasks (cf. Degen & Tanenhaus, 2015).

4.4. *Final thoughts*

Large-scale, systematic work on individual differences in pragmatic communication is only beginning. The present study is a first step toward incorporating EF and ToM measures into research on adults' pragmatic comprehension and could be extended to a broader range of pragmatic phenomena (see also Brown-Schmidt, 2009; Happé, 1993; Ryskin, Benjamin, Tullis, & Brown-Schmidt, 2015). Since neither ToM nor EF is a unitary construct, individual differences research in pragmatics needs to be based on a sound and theoretically motivated connection between the pragmatic phenomena themselves (within the specific tasks used to test them) and their specific ToM/EF presuppositions. This approach has the potential to specify the precise mechanisms underlying adults' processing of pragmatic meaning.

Acknowledgments

We thank Markus J. van Ackeren, Daniel Casasanto, Shirley-Ann Rueschemeyer, and Katarzyna Jankowiak for making the stimuli from Van Ackeren et al. (2012) and Jankowiak et al. (2017) available to us. This research was supported by NSF grant 1632849.

Notes

1. Other participant characteristics such as personality traits have also been proposed as sources of variation for scalar implicature computation (Heyman & Schaeken, 2015; Antoniou et al., 2016). However, the evidence for such proposals is weak at best (see Antoniou et al., 2016 for discussion); hence, we do not include them here.
2. A related strand of prior work has measured "social-communicative skills" using the Autism Quotient questionnaire (AQ; Baron-Cohen et al., 2001) and connected those skills to scalar implicature computation. The AQ was originally developed as a self-assessment tool for adults that could potentially be used to screen for autism spectrum conditions and measures the number of autistic traits a person possesses. On the Communicative Subscale of the AQ, which is often used in studies of pragmatics, examples of such autistic traits include being slow to understand a joke and having difficulties with turn-taking in telephone conversations. Prior studies have failed to find evidence of a link between AQ scores and under-informativeness judgments (Antoniou et al., 2016; Barbet & Thierry, 2016; Fairchild & Papafragou, 2018; Heyman & Schaeken, 2015), but other studies have reported that neural responses to under-informative statements differ between high-AQ and low-AQ groups (Nieuwland, Ditman, & Kuperberg, 2010; Zhaio, Liu, Chen, & Chen, 2015). Notice that the AQ is highly metacognitive: It does not directly measure an individual's ability to, for example, understand a joke or have a telephone conversation but requires participants to report on their own social skills. Furthermore, unlike ToM tasks, the AQ does not directly measure the ability to reason about others' mental states.
3. Hochstein et al. (2017) reported that their ASD participants gave mental state justifications 72% of the time on a version of Happé's (1994) Strange Stories task but did not include data on their control group of neurotypical adults. For comparison, using Hochstein et al.'s coding scheme, our adult participants in Experiment 1 gave mental state justifications 78% of the time.
4. To investigate the extent to which scalar implicature judgments are consistent within a single individual, a correlation analysis was performed between scores on the simple scalar implicature task (number of pragmatic responses out of 5) and mean Under-Informative sentence ratings in the dual scalar implicature task (all trials in the low Cognitive Load control condition). A Kendall's tau analysis was chosen to account for the positively skewed scores on the simple task, which represented the fact that people tended to respond to the pragmatically enriched meaning of the utterances. When participants were given a 5-point scale for responses, scores were fairly evenly distributed across the entire range of possible values. Despite these task differences in overall data patterns, scores on the two tasks were significantly negatively correlated, $\tau_b(176) = -0.292$, $p < .001$, suggesting a general tendency within an individual to respond either pragmatically or

logically. However, the correlation is only a moderate one, indicating that individuals do vary somewhat in the way they respond to the pragmatic meaning of an utterance across tasks (for similar observations, see Degen & Tanenhaus, 2015; Sikos, Kim, & Grodner, 2019; Tavano & Kaiser, 2010).

5. We repeated this analysis using ratings of Under-Informative sentences in the dual scalar implicature task (all trials) as the dependent variable and found the same results. The model was significant overall, $F(2, 175) = 9.878$, $p < .001$, and ToM ($\beta = -0.329$, $t = -3.429$, $p < .001$) but not EF ($\beta = -0.128$, $t = -1.386$, $p = .167$) was associated with sentence ratings.
6. To address the possibility that the linear regression analysis would not sufficiently handle potential clustering of scores at the scale boundaries, we reanalyzed the Pragmatic Score data in the simple scalar implicature task using a logistic analysis. We first transformed the Pragmatic Score into proportions representing the number of "Bad" ratings given by each participant in Under-Informative trials over the total number of Under-Informative trials per participant. A logistic model was then fitted to these pragmatic proportions with EF and ToM as predictors. The results replicated the findings of the multiple linear regression analysis. pragmatic proportions were significantly predicted by ToM ($p < 0.05$) but not EF ($p = 0.631$).
7. We repeated this analysis and the previous one involving Pragmatic Score using a single measure of EF (digit span performance). The results remained unchanged.
8. One could imagine alternative ways in which ToM might relate to SI performance; for instance, a hearer with advanced ToM skills might be more likely to adopt a *logical* interpretation of an under-informative utterance thinking that the speaker might not be knowledgeable. However, our data militate against this alternative possibility. In our SI task (and all similar tasks in the literature; see Katsos & Bishop, 2011), logical responding indicates either failure to draw a scalar implicature, or preference to disregard the implicature during responding in favor of the logical content.
9. This finding, along with evidence that even infants have some basic ToM abilities (e.g., Onishi & Baillargeon, 2005; Southgate, Senju, & Csibra, 2007; but see Kulke, Von Duhn, Schneider, & Rakoczy, 2018), suggests that EF is needed to complete the types of complex instances of mental-state reasoning that adults frequently engage in (and would have to engage in for pragmatic inferences), but may not be a necessity for simpler ToM processes.

References

Antoniou, K., Cummins, C., & Katsos, N. (2016). Why only some adults reject under-informative utterances. *Journal of Pragmatics*, *99*, 78–95.

Apperly, I. A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology*, *65*(5), 825–839.

Apperly, I. A., Back, E., Samson, D., & France, L. (2008). The cost of thinking about false beliefs: Evidence from adults' performance on a non-inferential theory of mind task. *Cognition*, *106*(3), 1093–1108.

Apperly, I. A., Samson, D., Chiavarino, C., & Humphreys, G. W. (2004). Frontal and temporo-parietal lobe contributions to theory of mind: Neuropsychological evidence from a false-belief task with reduced language and executive demands. *Journal of Cognitive Neuroscience*, *16*(10), 1773–1784.

Bambini, V., Ghio, M., Moro, A., & Schumacher, P. B. (2013). Differentiating among pragmatic uses of words through timed sensicality judgments. *Frontiers in Psychology*, *4*, 938.

Barbet, C., & Thierry, G. (2016). Some alternatives? Event-related potential investigation of literal and pragmatic interpretations of some presented in isolation. *Frontiers in Psychology*, *7*, 1479.

Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: Evidence from very high functioning adults with autism or Asperger syndrome. *Journal of Child Psychology and Psychiatry*, *38*(7), 813–822.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*, 37–46.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(2), 241–251.

Bergen, L., & Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(5), 1450–1460.

Bethell-Fox, C. E., & Shepard, R. N. (1988). Mental rotation: Effects of stimulus complexity and familiarity. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 12–23.

Bonnefon, J. F., Feeney, A., & Villejoubert, G. (2009). When some is actually all: Scalar inferences in face-threatening contexts. *Cognition*, *112*(2), 249–258.

Bott, L., Bailey, T. M., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language*, *66*(1), 123–142.

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

Bowdle, B. F., & Gentner, D. (2005). The career of metaphor. *Psychological Review*, *112*(1), 193–216.

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

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

Brown-Schmidt, S. (2009). The role of executive function in perspective taking during online language comprehension. *Psychonomic Bulletin & Review*, *16*(5), 893–900.

Bull, R., Phillips, L. H., & Conway, C. A. (2008). The role of control functions in mentalizing: Dual-task studies of theory of mind and executive function. *Cognition*, *107*(2), 663–672.

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development*, *72*(4), 1032–1053.

Carlson, S. M., Moses, L. J., & Breton, C. (2002). How specific is the relation between executive function and theory of mind? Contributions of inhibitory control and working memory. *Infant and Child Development*, *11*(2), 73–92.

Carston, R. (1995). Quantity maxims and generalised implicature. *Lingua*, *96*(4), 213–244.

Champagne-Lavau, M., & Joanette, Y. (2009). Pragmatics, theory of mind and executive functions after a right-hemisphere lesion: Different patterns of deficits. *Journal of Neurolinguistics*, *22*(5), 413–426.

Chevallier, C., Wilson, D., Happé, F., & Noveck, I. (2010). Scalar inferences in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *40*(9), 1104–1117.

Chiappe, D. L., & Chiappe, P. (2007). The role of working memory in metaphor production and comprehension. *Journal of Memory and Language*, *56*(2), 172–188.

Cooper, R. P. (2010). Cognitive control: Componential or emergent? *Topics in Cognitive Science*, *2*(4), 598–613.

Coulson, S., & Lovett, C. (2010). Comprehension of non-conventional indirect requests: An event-related brain potential study. *Italian Journal of Linguistics*, *22*(1), 107–124.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*(11), 671–684.

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology*, *54*(2), 128–133.

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

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168.

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

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, *45*(3), 259–268.

Dumontheil, I., Apperly, I. A., & Blakemore, S. J. (2010). Online usage of theory of mind continues to develop in late adolescence. *Developmental Science*, *13*(2), 331–338.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*(3), 309–331.

Epley, N., Morewedge, C. K., & Keysar, B. (2004). Perspective taking in children and adults: Equivalent egocentrism but differential correction. *Journal of Experimental Social Psychology*, *40*(6), 760–768.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149.

Eviatar, Z., & Just, M. A. (2006). Brain correlates of discourse processing: An fMRI investigation of irony and conventional metaphor comprehension. *Neuropsychologia*, *44*(12), 2348–2359.

Fairchild, S., Mathis, A., & Papafragou, A. (2020). Pragmatics and social meaning: Understanding under-informativeness in native and non-native speakers. *Cognition*, *200*, 104171.

Fairchild, S., & Papafragou, A. (2018). Sins of omission are more likely to be forgiven in non-native speakers. *Cognition*, *181*, 80–92.

Feng, S., Ye, X., Mao, L., & Yue, X. (2014). The activation of theory of mind network differentiates between point-to-self and point-to-other verbal jokes: An fMRI study. *Neuroscience Letters*, *564*, 32–36.

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development*, *59*(1), 47–59.

Geurts, B. (2010). *Quantity implicatures*. Cambridge, UK: Cambridge University Press.

Glucksberg, S. (2003). The psycholinguistics of metaphor. *Trends in Cognitive Sciences*, *7*(2), 92–96.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (pp. 41–58). New York: Academic Press.

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

Guasti, M. T., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., & Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes*, *20*(5), 667–696.

Happé, F. G. (1993). Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition*, *48*(2), 101–119.

Happé, F. G. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, *24*(2), 129–154.

Heyman, T., & Schaeken, W. (2015). Some differences in some: Examining variability in the interpretation of scalars using latent class analysis. *Psychologica Belgica*, *55*(1), 1–18.

Hirschberg, J. L. B. (1985). *A theory of scalar implicature*. Philadelphia: University of Pennsylvania.

Hochstein, L., Bale, A., & Barner, D. (2017). Scalar implicature in absence of epistemic reasoning? The case of autism spectrum disorder. *Language Learning and Development*, *14*(3), 1–17.

Hochstein, L., Bale, A., Fox, D., & Barner, D. (2014). Ignorance and inference: Do problems with Gricean epistemic reasoning explain children's difficulty with scalar implicature? *Journal of Semantics*, *33*(1), 107–135.

Horn, L. R. (1972). On the semantic properties of the logical operators in English. Doctoral dissertations, UCLA.

Horn, L. R.(1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Georgetown University round table on languages and linguistics* (pp. 11–42). Washington, DC: Georgetown University Press.

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

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

Hughes, C., & Ensor, R. (2007). Executive function and theory of mind: Predictive relations from ages 2 to 4. *Developmental Psychology*, *43*(6), 1447–1456.

Jankowiak, K., Rataj, K., & Naskręcki, R. (2017). To electrify bilingualism: Electrophysiological insights into bilingual metaphor comprehension. *PLoS ONE*, *12*(4), e0175578.

Jolliffe, T., & Baron-Cohen, S. (1999). The strange stories test: A replication with high-functioning adults with autism or Asperger syndrome. *Journal of Autism and Developmental Disorders*, *29*(5), 395–406.

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, *9*(4), 637–671.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189–217.

Katsos, N., & Bishop, D. V. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, *120*(1), 67–81.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32–38.

Kulke, L., von Duhn, B., Schneider, D., & Rakoczy, H. (2018). Is implicit theory of mind a real and robust phenomenon? Results from a systematic replication study. *Psychological Science*, *29*(6), 888–900.

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

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*(3), 551–556.

Loukusa, S., & Moilanen, I. (2009). Pragmatic inference abilities in individuals with Asperger syndrome or high-functioning autism. A review. *Research in Autism Spectrum Disorders*, *3*(4), 890–904.

Marty, P. P., & Chemla, E. (2013). Scalar implicatures: Working memory and a comparison with only. *Frontiers in Psychology*, *4*, 403–412.

Mashal, N. (2013). The role of working memory in the comprehension of unfamiliar and familiar metaphors. *Language and Cognition*, *5*(4), 409–436.

Matthews, D., Biney, H., & Abbot-Smith, K. (2018). Individual differences in children's pragmatic ability: A review of associations with formal language, social cognition and executive functions. *Language Learning and Development*, *14*(3), 186–223.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100.

Newschaffer, C. J., Croen, L. A., Daniels, J., Giarelli, E., Grether, J. K., Levy, S. E., Mandell, D. S., Miller, L. A., Pinto-Martin, J., Reaven, J., Reynolds, A. M., Rice, C. E., Schendel, D., & Reynolds, A. M. (2007). The epidemiology of autism spectrum disorders. *Annual Review of Public Health*, *28*, 235–258.

Nieuwland, M. S., Ditman, T., & Kuperberg, G. R. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language*, *63* (3), 324–346.

Norbury, C. F. (2005). The relationship between theory of mind and metaphor: Evidence from children with language impairment and autistic spectrum disorder. *British Journal of Developmental Psychology*, *23*(3), 383–399.

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, *78*(2), 165–188.

Noveck, I. A., Guelminger, R., Georgieff, N., & Labruyere, N. (2007). What autism can reveal about every… not sentences. *Journal of Semantics*, *24*(1), 73–90.

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

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308* (5719), 255–258.

Papafragou, A., Friedberg, C., & Cohen, M. L. (2018). The role of speaker knowledge in children's pragmatic inferences. *Child Development*, *89*, 1642–1656.

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: Experiments at the semantics–pragmatics interface. *Cognition*, *86*(3), 253–282.

Pijnacker, J., Hagoort, P., Buitelaar, J., Teunisse, J. P., & Geurts, B. (2009). Pragmatic inferences in high-functioning adults with autism and Asperger syndrome. *Journal of Autism and Developmental Disorders*, *39*(4), 607–618.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2017). *nlme: Linear and nonlinear mixed effects models*. R package version 3.1-131. Retrieved from https://CRAN.R-project.org/package=nlme

Politzer-Ahles, S., Fiorentino, R., Jiang, X., & Zhou, X. (2013). Distinct neural correlates for pragmatic and semantic meaning processing: An event-related potential investigation of scalar implicature processing using picture-sentence verification. *Brain Research*, *1490*, 134–152.

Politzer-Ahles, S., & Gwilliams, L. (2015). Involvement of prefrontal cortex in scalar implicatures: Evidence from magnetoencephalography. *Language, Cognition and Neuroscience*, *30*(7), 853–866.

Pouscoulous, N., Noveck, I. A., Politzer, G., & Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Language Acquisition*, *14*(4), 347–375.

Prat, C. S., Mason, R. A., & Just, M. A. (2012). An fMRI investigation of analogical mapping in metaphor comprehension: The influence of context and individual cognitive capacities on processing demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 282–294.

Qureshi, A. W., Apperly, I. A., & Samson, D. (2010). Executive function is necessary for perspective selection, not Level-1 visual perspective calculation: Evidence from a dual-task study of adults. *Cognition*, *117*(2), 230–236.

R Core Team. (2015). *R: A language and environment for statistical computing*, Vienna, Austria: R Foundation for Statistical Computing.

Ryskin, R. A., Benjamin, A. S., Tullis, J., & Brown-Schmidt, S. (2015). Perspective-taking in comprehension, production, and memory: An individual differences approach. *Journal of Experimental Psychology: General*, *144*(5), 898–915.

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(5), 1255–1266.

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy*, *27*(3), 367–391.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *NeuroImage*, *19*(4), 1835–1842.

Schaeken, W., Van Haeren, M., & Bambini, V. (2018). The understanding of scalar implicatures in children with autism spectrum disorder: Dichotomized responses to violations of informativeness. *Frontiers in Psychology*, *9*, 1266.

Shetreet, E., Chierchia, G., & Gaab, N. (2014). When some is not every: Dissociating scalar implicature generation and mismatch. *Human Brain Mapping*, *35*(4), 1503–1514.

Sikos, L., Kim, M., & Grodner, D. (2019). Social context modulates tolerance for pragmatic violations in binary but not graded judgments. *Frontiers in Psychology*, *10*, 510.

Skordos, D., & Papafragou, A. (2016). Children's derivation of scalar implicatures: Alternatives and relevance. *Cognition*, *153*, 6–18.

Slabakova, R. (2010). Scalar implicatures in second language acquisition. *Lingua*, *120*(10), 2444–2462.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*(7), 587–592.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*, Cambridge, MA: Blackwell.

Spotorno, N., Koun, E., Prado, J., Van Der Henst, J. B., & Noveck, I. A. (2012). Neural evidence that utterance-processing entails mentalizing: The case of irony. *NeuroImage*, *63*(1), 25–39.

Tavano, E., & Kaiser, E. (2010). Processing scalar implicature: What can individual differences tell us?, University of Pennsylvania Working Papers in Linguistics. *16*(1), 214–224.

Van Ackeren, M. J., Casasanto, D., Bekkering, H., Hagoort, P., & Rueschemeyer, S. A. (2012). Pragmatics in action: Indirect requests engage theory of mind areas and the cortical motor network. *Journal of Cognitive Neuroscience*, *24*(11), 2237–2247.

Wampers, M., Schrauwen, S., De Hert, M., Gielen, L., & Schaeken, W. (2018). Patients with psychosis struggle with scalar implicatures. *Schizophrenia Research*, *195*, 97–102.

Waters, G. S., & Caplan, D. (2003). The reliability and stability of verbal working memory measures. *Behavior Research Methods, Instruments, & Computers*, *35*(4), 550–564.

Wechsler, D. (1944). *The measurement of adult intelligence*. Baltimore, MD: Williams & Wilkins.

Wechsler, D. (1981). *WAIS-R manual: Wechsler adult intelligence scale-revised*, San Antonio, TX: Psychological Corporation.

Winner, E., Brownell, H., Happé, F., Blum, A., & Pincus, D. (1998). Distinguishing lies from jokes: Theory of mind deficits and discourse interpretation in right hemisphere brain-damaged patients. *Brain and Language*, *62*(1), 89–106.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Appendix A: Stimuli used in the dual scalar implicature task in Experiment 1.

Appendix B: Stimuli used in the simple scalar implicature task in Experiments 1 and 2 (from Bott & Noveck, 2004).

Appendix C: Stimuli used in the metaphor task in Experiment 2 (from Jankowiak, Rataj, & Naskręcki, 2017).

Appendix D: Stimuli used in the indirect request task in Experiment 2 (adapted from Van Ackeren, Casasanto, Bekkering, Hagoort, & Rueschemeyer, 2012).