Why second-language speakers sometimes but not always derive scalar inferences like first-language speakers: Effects of task demands

Ahmed Khorsheed

Universiti Putri Malaysia

Bob van Tiel

Radboud University Nijmegen

Author note

We declare no conflict of interest.

To appear in: *Language Acquisition*
https://doi.org/10.1080/10489223.2024.2383574

Abstract

Many studies report that the computation of scalar inferences, such as the inference from ‘some’ to ‘not all’, is cognitively costly. However, a number of studies on scalar inferences in L2 suggest the opposite conclusion, reporting that, even though processing L2-input is often thought to be cognitively demanding, people are at least as likely to compute scalar inferences in L2 as L1. Here, we propose that cognitive difficulties with computing scalar inferences in L2 may be masked by the use of experimental tasks that encourage participants to take their time to process target sentences. Thus, we found that, in a self-paced pen-and-paper experiment, participants did not significantly differ in the propensity with which they computed scalar inferences in L1 and L2. However, in a computerised experiment in which sentences were flashed on screen, low-proficiency L2 speakers were significantly less likely to compute scalar inferences in L2 than L1. These results underscore the importance of analysing task demands in research on pragmatics and multilingualism.

Why second-language speakers sometimes but not always derive scalar inferences like first-language speakers: Effects of task demands

When we use language, what we mean to say often goes beyond the literal meaning of the words that we produce. To illustrate, consider the dialogue in (1) (taken from Carston, 2004).

(1) A: Did the children's summer camp go well?

B: Some of them got stomach flu.

(i) The summer camp didn't go as well as hoped.

(ii) Not all the children got stomach flu.

B's answer does not directly address the question that A asked. However, based on the assumption that B is *cooperative*, i.e., does their best to further the goal of the conversation, it may be inferred that the summer camp did not go as well as hoped. Moreover, again based on the assumption that B is cooperative, it may be inferred that not all of the children got stomach flu, since, if that were the case, it would have been informative—and hence cooperative—for B to have said so. Grice (1975) introduced the term *conversational implicature* to refer to such inferences that can be justified on the basis of what is literally said and the assumption that the speaker is cooperative.

Intuitively, there is an important distinction between the two conversational implicatures associated with B's answer. The inference that the summer camp did not go as well as planned is heavily reliant on the particular context created by A's utterance. If B's answer had addressed a different question (e.g., 'What happened to the children?'), nothing about the summer camp would have been inferred. By contradistinction, the inference that not all of the children got stomach flu is relatively impervious to contextual variation. Grice called

these two types of inferences *particularised* and *generalised* conversational implicatures, respectively.

In the literature, the kind of generalised conversational implicature exemplified by the inference from 'some' to 'not all' has become known as a *scalar inference*. This name stems from the idea that words like 'some' are associated with lexical scales. In the case at hand, the relevant scale is <some, all>. By using a weaker term on the scale (e.g., 'some'), the speaker may imply that the corresponding sentence with the stronger term (e.g., 'all') is false (e.g., Gazdar, 1979; Geurts, 2010; Horn, 1972).

Grice's theory of conversational implicature was aimed at providing a rational justification of the pragmatic inferences that we derive, rather than a description of the psychological processes that underlie their computation (Geurts & Rubio-Fernández, 2015; van Tiel & Schaeken, 2017). However, Grice's theory has inspired several authors to provide such psychological theories. Perhaps the two most prominent examples are *relevance theory* (Sperber & Wilson, 1995) and *defaultism* (Levinson, 2000). These two theories hold diametrically opposite views on the cognitive mechanism responsible for implicature processing, particularly in the case of scalar inferences.

According to relevance theory, human communication is geared towards the maximisation of relevance, where relevance is defined as the degree to which new information makes "a worthwhile difference" to one's representation of the world (Sperber & Wilson, 2006, p. 608). Consequently, according to relevance theory, the hearer should only compute scalar inferences if these are sufficiently relevant to them, and if the processing effort that is needed for their computation is justified by the expected interpretative benefits. In out-of-the-blue contexts—such as those used in most experimental studies—the computation of scalar inferences is assumed to be cognitively costly, although the size of this cost is variable and

dependent on the effort needed for bridging the gap between the literal meaning and the pragmatically enriched meaning.

By contrast, Levinson (2000) explicitly denies that the computation of scalar inferences is cognitively costly. Indeed, Levinson argues that the pragmatically enriched meaning should be *easier* to retrieve than the literal meaning. In particular, Levinson proposes that scalar inferences are default inferences, i.e., they are computed automatically unless there are compelling reasons not to. Levinson conceptually motivates his defaultist position by observing that scalar inferences are ubiquitous in natural language. Hence, he argues, it would be inefficient if hearers had to go through a protracted reasoning process to compute scalar inferences whenever they encountered a scalar expression. According to Levinson, since language has evolved for communicative efficiency, scalar inferences have become "hard-wired" into the meanings of scalar expressions like 'some'.

Recent proposals have tried to find middle ground between relevance theory and Levinson's defaultism. These recent proposals argue that the presence or absence of a processing cost for scalar inferences depends on various factors, such as the question under discussion (e.g., Ronai & Xiang, 2021), the structural characteristics of the alternatives (e.g., Chemla & Bott, 2014; van Tiel & Schaeken, 2017), the naturalness of the utterance (Degen & Tanenhaus, 2015), and the polarity of the scalar expression (e.g. van Tiel & Pankratz, 2021; van Tiel et al., 2019). See Khorsheed, Price, and van Tiel (2022) for an overview.

In one of the first studies that tried to adjudicate between relevance theory and Levinson's defaultism, Bott and Noveck (2004) carried out a series of truth-value judgement tasks in which participants evaluated the truth value of categorical sentences with 'some' and 'all', such as:

(2) Some elephants are mammals.

(i) At least some elephants are mammals. *Literal*

(ii) Some but not all elephants are mammals. *Pragmatic*

On its *literal* interpretation, (2) merely says that there are elephants that are mammals. However, this sentence is associated with a scalar inference according to which not all elephants are mammals. If we take the literal interpretation and the scalar inference together, we arrive at the *pragmatic* interpretation according to which at least some, but not all, elephants are mammals. Hence, on their literal interpretation, sentences such as (2) are true; on their pragmatic interpretation, they are false.

In Bott and Noveck's Exp. 3, participants could freely provide their responses. Here, Bott and Noveck found that participants were roughly equally likely to accept or reject underinformative sentences such as (2). However, in terms of response times, they found that 'false' responses took significantly longer than 'true' responses. This difference in response times was absent in a control condition with sentences that were unambiguously true (3a) or false (3b).

(3) a. Some mammals are elephants.

b. Some elephants are insects.

Based on these findings, Bott and Noveck concluded that pragmatic interpretations require more time and deeper processing than literal interpretations. This conclusion aligns with relevance theory, but counters Levinson's assertion that the pragmatic interpretation is the default one.

The conclusion that the computation of scalar inferences is cognitively effortful has been confirmed in numerous other studies using a wide array of tasks (e.g., Breheny et al.,

2006; De Neys & Schaeken, 2007; Huang & Snedeker, 2018; Marty et al., 2013; Tomlinson Jr. et al., 2013, but cf. Grodner et al., 2010). Moreover, the cognitive cost of scalar inferences was also observed for some—though not all—scalar expressions other than 'some' (e.g., Romoli & Schwarz, 2015; van Tiel & Pankratz, 2021; van Tiel et al., 2019), as well as for numerals, which are sometimes said to trigger scalar inferences (e.g., Noveck et al., 2022; Spychalska et al., 2016). Moreover, the cognitive cost was observed for individuals that vary in terms of their age, personality traits, and clinical profile (e.g., Janssens et al., 2014; Schaeken et al., 2018; Skordos & Papafragou, 2016). In summary, there is substantive evidence that the computation of scalar inferences is cognitively costly.

Given this evidence, it is natural to expect that people are less likely to compute scalar inferences in their second language (L2) compared to their first (L1). After all, processing L2 input is known to be cognitively demanding, which limits the availability of the cognitive resources needed for computing scalar inferences (e.g., Clahsen & Felser, 2006; Green, 1986; 1998; Juffs, 2001; White & Juffs, 1998). In stark contrast to this prediction, many existing studies on the topic report that people are equally, and in some cases even *more* likely to compute scalar inferences in L2 than L1 (e.g., Dupuy et al., 2019; Feng & Cho, 2019; Lin, 2016; Slabakova, 2010; Snape & Hosoi, 2018).

The goal of our study is to offer an explanation for this apparent incongruity. We will argue that, in fact, people do experience difficulties computing scalar inferences in L2, but that these difficulties may be masked by the use of experimental tasks in which participants can take their time to process and evaluate the relevant sentences. By means of two experiments, we will show that, when participants can process and evaluate sentences at their leisure, they do not significantly differ in the propensity with which they compute scalar inferences in L1 and L2. But if we limit participants' access to the sentences by flashing each word on the screen for a brief amount of time, low-proficiency speakers become less likely to compute scalar

inferences in L2 than in L1. We suggest that low-proficiency speakers may use metalinguistic reasoning to compute scalar inferences in their L2 but not—or at least to a significantly lesser extent—in their L1.

In the next section, we provide a brief overview of previous studies on scalar inferences in L2. Afterwards, we describe our experiments.

Scalar inferences in L2

While there is abundant data on the processing of scalar inferences in L1 (see Noveck, 2018, for an overview), data on the processing of scalar inferences in L2 is comparatively scarce and currently poorly understood (Holtgraves et al., 2019).

Slabakova (2010) was the first to directly study the derivation of scalar inferences in L2. She tested three groups of participants: (i) native speakers of English, (ii) native speakers of Korean, and (iii) native speakers of Korean who spoke English as a second language. The first two groups were tested in their L1; the last one in their L2.

In her Exp. 1a, Slabakova carried out a sentence verification task similar to the one used by Bott and Noveck (2004) (see above for discussion). Slabakova found that the Korean-speaking learners of English were more likely to disagree with underinformative sentences with 'some'—such as (2) above—in their L2 when compared to native speakers of English or Korean in their L1.

Slabakova took this finding to support Levinson's defaultist account, since it seems to show that scalar inferences are computed automatically, and that people who process L2 input have fewer cognitive resources at their disposal to overturn these default inferences.

Although Slabakova's results were partly replicated by Snape and Hosoi (2018), other studies report markedly different results. A case in point is the study by Dupuy and colleagues (2019). Dupuy and colleagues tested two groups of participants: (i) monolingual speakers of

French, and (ii) native speakers of French who spoke English as a second language. Participants were tested using a sentence-picture verification task, in which they were presented with sentences such as:

(4) The boy has hidden some of the cars.

In the target condition, the corresponding picture made it clear that the boy had, in fact, hidden all of the cars, thus verifying the literal meaning of the sentence but falsifying its scalar inference. Analogously to Slabakova's sentence verification task, participants were thus expected to reject the sentence if they computed the scalar inference.

In their Exp. 1a, bilingual participants were tested both in their L1 and L2. In line with Slabakova, it was found that participants were more likely to reject underinformative sentences with 'some', such as (4), in their L2, when compared to monolingual French speakers in their L1. However, strikingly, Dupuy and colleagues found a similar difference when comparing the bilinguals' responses in their L1 with monolinguals' responses in their L1. Dupuy and colleagues thus found an overall facilitatory effect of bilingualism rather than an effect of language proficiency per se.

Dupuy and colleagues hypothesised that this effect of bilingualism was a product of testing bilingual participants in both their L1 and L2 consecutively. To test this explanation, their Exp. 1b replicated Exp. 1a, but tested bilinguals only in their L2. In line with their predictions, it was found that, in this experiment, L2-learners did not outperform monolingual speakers in providing pragmatic responses. Hence, it seems that, by announcing that bilingual participants in the first experiment would be tested in both of their languages, Dupuy and colleagues primed them to compute scalar inferences, e.g., because this announcement made

participants more self-conscious of their linguistic behaviour, which led them to think more deeply about the sentences that they were presented with.

Whereas Slabakova found that people were more likely to compute scalar inferences in L2 than L1, Dupuy and colleagues thus ultimately found no difference between L1 and L2. Most studies on the topic confirm either of these two patterns of results (e.g., Antoniou & Katsos, 2017; Antoniou et al., 2018; Feng & Cho, 2019; Lin, 2016; Snape & Hosoi, 2018). However, in a more recent study, Mazzaggio and colleagues (2021) found that people were *less* likely to compute scalar inferences in L2 than L1.

In their Exp. 1, Mazzaggio and colleagues tested native Italian speakers who spoke English as a second language. These participants were divided into two groups: the first group was tested in their L1; the second in their L2. Mazzaggio and colleagues carried out a sentence verification task that was similar to Slabakova's Exp. 1a. However, there were two important methodological differences. First, Mazzaggio and colleagues presented the sentences auditorily instead of in written form. Second, participants had to respond (i.e., indicate if the sentence was correct or incorrect) within a time frame of three seconds.

In contrast with much of the preceding literature, Mazzaggio and colleagues found that participants were significantly less likely to compute scalar inferences in L2 than in L1. To explain this apparently anomalous finding, they point to the two methodological differences between their experiment and that of, e.g., Slabakova. One of the consequences of these methodological differences was that participants could no longer take their time to process and evaluate the sentences, since these were presented auditorily rather than visually, and since participants had to respond within three seconds.

In order to explain why this manipulation differentially affected the rates of scalar inferences in L1 and L2, we might hypothesise that there are two ways of computing scalar inferences. First, scalar inferences can be "grasped intuitively" (Grice, 1975, p. 50) by

identifying the relevant regularity, e.g., the regularity that ‘some’ tends to imply ‘not all’. Second, scalar inferences can be “worked out” (ibid.) by reasoning about the speaker’s intended meaning on the basis of what they said and the assumption that they are cooperative.

The first, intuitive route requires knowledge of language-specific pragmatic regularities, whereas the second, effortful route only requires language-universal knowledge of pragmatic principles. Hence, L1 speakers may use the first route, while L2 speakers have to rely on the second route because they are not adequately familiar with the pragmatic regularities in their L2. As a consequence, motivated participants may be equally or even more likely to compute scalar inferences in L2 than L1 as long as they have the time and cognitive resources needed for going through the pragmatic reasoning process.

A provocative piece of evidence in favour of this explanation comes from Khorsheed, Rashid, and colleagues (2022), who carried out an experiment in which they tested a large group of native speakers of Malay who spoke English as a second language. These participants were tested in their L2, i.e., English. Based on a local proficiency test, participants were divided into low-proficiency and high-proficiency speakers of English. Khorsheed, Rashid, and colleagues carried out a sentence verification task that was similar to Slabakova’s Exp. 1a.

First, Khorsheed, Rashid, and colleagues found that high-proficiency speakers were significantly more likely to reject underinformative sentences with ‘some’ than low-proficiency speakers. This observation is concordant with the results reported by Mazzaggio and colleagues, and suggests that, when people require more cognitive resources for processing linguistic input—as low-proficiency speakers do in comparison to high-proficiency speakers—they become less likely to compute scalar inferences. Indeed, as in the study by Mazzaggio and colleagues, participants in the study of Khorsheed, Rashid, and colleagues did not have unlimited access to the test sentences because these were presented by flashing each word on the screen for a brief amount of time (similarly to the experiments of Bott & Noveck, 2004).

Khorsheed, Rashid, and colleagues also measured participants' response times. Similarly to Bott and Noveck, they found that participants were significantly slower when they computed the scalar inference (i.e., when they answered 'false' in the target condition) than when they did not (i.e., when they answered 'true'). Critically, Khorsheed, Rashid, and colleagues found that this delay in response times was more pronounced for low-proficiency participants than for high-proficiency participants. This observation confirms the idea that less proficient speakers are more likely to rely on effortful processing to compute scalar inferences.

While the results reported by Mazzaggio, Khorsheed, and their colleagues provide tantalising evidence that less proficient speakers rely on effortful processing for computing scalar inferences, their studies have an important limitation: they did not compare their results directly to an experimental setting in which participants' cognitive resources were *not* burdened; only to results from previous studies that tested different linguistic profiles using different experimental tasks. Given that previous results were inconsistent, it is important to test the proposed explanation on the basis of a direct comparison between experiments that differ in whether they allow participants to engage in metalinguistic reasoning about the target sentences.

In summary, the current experimental record calls for a more systematic investigation into the role of task demands in the study of scalar inferences, so as to test the hypothesis that less proficient speakers rely on effortful processing for computing scalar inferences. In the next section, we provide an overview of the experiments that we conducted to test this hypothesis.

Our experiments

We carried out two experiments to determine whether the probability of computing scalar inferences in L2 is influenced by whether participants can take their time to process and evaluate sentences. In both experiments, we tested native speakers of Malay who spoke English

as a second language. Following Khorsheed, Rashid, and colleagues (2022), we divided these participants into high-proficiency and low-proficiency speakers of English.

Both experiments used the same experimental task as Bott and Noveck (2004, Exp. 3) (which was also used in many of the L2 studies described above). In this task, participants had to indicate whether they agreed or disagreed with sentences with 'some' and 'all'. The target condition consisted of sentences such as (5) which were literally true but carried a scalar inference that was false.

(5) Some parrots are birds.

Exp. 1 was a pen-and-paper task in which participants could take as long as they wanted to read and evaluate the sentences, and, as a consequence, to compute the scalar inference. By contrast, Exp. 2 mirrored Bott and Noveck's experiment in that sentences were presented by flashing each word on screen for a short period of time. Moreover, in Exp. 2 but not Exp. 1, participants' response times were measured, and they were instructed to respond as quickly as possible. Exp. 2 was thus a replication of the study of Khorsheed, Rashid, and colleagues, except that we also tested participants in their L1.

Our study differed from Slabakova (2010) in that all of our participants were bilingual. Hence, in line with Dupuy and colleagues (2019), we expect that, in Exp. 1, we do not find a significant difference in the propensity with which participants compute scalar inferences in L1 and L2. However, we predict that, in Exp. 2, low-proficiency English speakers are less likely to compute scalar inferences in their L2, given their reliance on effortful pragmatic reasoning to derive these inferences. Following the work of Khorsheed, Rashid, and colleagues, we further hypothesise that this effortful derivation process is reflected in response times, i.e.,

that low-proficiency speakers take longer to compute scalar inferences in their L2 in comparison with high-proficiency speakers.

Finally, in a more exploratory analysis, we studied the effect of working memory capacity on the computation of scalar inferences. If, as we suppose, low-proficiency speakers draw on effortful pragmatic reasoning to compute scalar inferences, we might expect that participants with fewer working memory resources will be less likely to go through this reasoning process.

Previous studies have investigated the effect of working memory capacity on the computation of scalar inferences in L1, though with varying results. On the one hand, Antoniou and colleagues (2016) and Yang and colleagues (2018) found that participants with more cognitive resources—including greater working memory capacity—were more likely to compute scalar inferences. On the other hand, Janssens and colleagues (2014) and Schaeken and colleagues (2021) found no significant effects of working memory capacity on scalar inference computation in children and patients with schizophrenia, respectively. Dieussaert and colleagues (2011) also found no overall effect of working memory capacity, though they found that it modulates effects of working memory load manipulations.

Lastly, Lin (2016) found that participants with a higher working memory capacity were more likely to accept the literal reading of sentences with ‘some’ in their L2. At the same time, Lin found that participants with a higher working memory capacity were also *more* likely to accept the pragmatic reading of the sentences with ‘some’. In addition, Lin appears to support his conclusions with inappropriate statistical analysis, using chi-squared tests on data from a within-participants design. Hence, Lin’s results should be interpreted with caution.

Given these contradictory findings, we consider our investigation of the effects of working memory capacity on scalar inferences in L2 as an exploratory analysis.

In the next section, we describe the two experiments and discuss the results.

Experiment 1: Pen-and-paper task

Participants

105 participants took part in this study. All participants were undergraduate students at Universiti Putra Malaysia. All but five participants were female. Their mean age was 22 (standard deviation: 2, range: 18–30).

Participants were divided into two main groups, depending on whether they were tested in their first language (L1-Malay) or second language (L2-English). The L1-Malay group comprised 35 participants. The L2-English group comprised 70 participants, including 33 with low English proficiency and 37 with high English proficiency. Participants' proficiency in English was determined by the Malaysian University English Test (MUET), which is used for university admissions in Malaysia. The MUET evaluates students' listening, speaking, reading, and writing skills and assigns them a band score ranging from 1 (very limited user) to 6 (highly proficient user). In our sample, 3 participants were classified in band 2, 30 in band 3, 31 in band 4, and 6 in band 5. Given the relatively small number of participants in bands 2 and 5, we combined participants from bands 2 and 3 into a low-proficiency group, and participants from bands 4 and 5 into a high-proficiency group.

Materials and procedure

Similarly to Bott and Noveck (2004), we carried out a truth-value judgement task in which participants were presented with categorical sentences with 'some' and 'all'. There were six types of test sentences, which were labelled T1 to T6. Table 1 shows an example sentence for each sentence type. Participants received the test in the form of a printed paper. They were instructed to read each sentence and judge whether, according to their intuitions, they were true or false. They could spend as much or as little time as they wanted to read and evaluate the

sentences. At the end of the experiment, each participant was given a credit worth $1 for their participation.

<Insert Table 1 about here>

One of our reviewers was concerned that we followed Bott and Noveck's (2004) original study in using the response option labels 'true' and 'false' rather than, e.g., 'right' and 'wrong'. Indeed, participants might hesitate to judge an underinformative sentence with 'some' as altogether false, given that its literal meaning is true (e.g., Jasbi et al., 2019; Katsos & Bishop, 2011). To our knowledge, potential effects of response option labels have never been directly tested. However, if any such effects are additive (e.g., if participants are overall more likely to judge underinformative sentences with 'some' as being wrong rather than false), we expect the same pattern of results for other choices of response option labels.

To create our test sentences, we employed a Latin square design, which involved combining an exemplar name (such as 'eagles') with a category name (such as 'birds'). This process resulted in a list of 54 sentences. For the two L2-English groups, the test sentences were presented in English, whereas for the L1-Malay group, they were presented in Malay. Sentence type T1 was the target condition, because sentences of this type are literally true but give rise to scalar inferences that are false. Hence, we expect that participants who compute the scalar inference will judge the sentence to be false. Sentence types T2 to T6 served as control sentences in the design.

To ensure that there were no semantic differences between the Malay and English test sentences, we conducted a rating test in which participants were asked to review pairs of English-Malay sentences and indicate, on a ten-point Likert scale, to what extent the test sentences in Malay and English expressed the same semantic meaning. A score of 1 indicated

the sentences held very different meanings, while 10 meant they had exactly the same meaning. For this test, we recruited 16 participants who were not part of the main study. Each participant was presented with eight pairs of Malay and English test sentences, comprising two filler pairs that were not equivalent in their semantic meaning, three target pairs containing the quantifiers 'some' and its Malay equivalent 'sesetengah', and three target pairs with the quantifiers 'all' and its Malay equivalent 'semua'.

For data analysis, the participants who did not exhibit full accuracy on the control items were removed from the data. This resulted in the removal of three participants. Following this, the mean ratings for the target English-Malay sentence pairs were computed. The results indicated that the target sentences were perceived as highly similar in their semantic meaning. More specifically, 'some' and 'all' in English were rated as semantically equivalent to their Malay counterparts 'sesetengah' and 'semua' (mean semantic similarity scores of 9.79 and 9.87, respectively). For more relevant discussion of the semantic and syntactic status of English and Malay quantifiers, see Khorsheed, Rashid, and colleagues (2022, p. 20).

Data treatment

Seven participants were removed from the analysis because they made mistakes in more than 20% of the control items (three from the Malay-L1 group, one from the high-proficiency English-L2 group, and three from the low-proficiency English-L2 group).

Results

Fig. 1 summarises the percentages of 'true' responses for each sentence type in each of the three language groups (L1-Malay, high-proficiency L2-English, and low-proficiency L2-English). Visually, there seem to be few differences across the three groups. The percentages

of 'true' responses in the target condition (i.e., sentence type T1) were 15%, 21%, and 22%, respectively.

<Insert Figure 1 about here>

First, we analysed whether the three language groups differed in the control conditions (i.e., sentence types T2–T6). All of the following analyses made use of the 'lme4' package (Bates, Mächler, Bolker, & Walker, 2015) in R (R Core Team, 2023). Thus, we constructed a binomial generalised linear mixed-effects model predicting responses (correct or incorrect) on the basis of condition (T2–T6), language group (L1-Malay, high-proficiency L2-English, low-proficiency L2-English), and their interaction, including random intercepts for participants and items. The effects of the fixed factors were estimated using model comparison with more parsimonious models using the 'anova' function. In all of the following analyses, categorical predictors were sum-coded for interpretability of the parameter estimates.

These analyses indicated that there was a significant effect of language group ($\chi^2(2) = 13.4$, $p = .001$). Neither of the other effects were significant (condition: $\chi^2(4) = 2.3$, $p = .67$; condition × language group: $\chi^2(8) = 7.8$, $p = .45$). Focusing on a model predicting responses on the basis of language group, we pairwise compared the language groups to determine which of them differed in terms of the proportion of correct responses. For this analysis, we relied on the 'multcomp' package (Hothorn, Bretz, & Westfall, 2008). It was found that the proportion of correct responses was significantly lower in the low-proficiency L2-English group (92%) compared to the high-proficiency L2-English group (96%, $\beta = -0.9$, $SE = 0.3$, $Z = -3.4$, $p = .002$), and compared to the L1-Malay group (96%, $\beta = 0.8$, $SE = 0.3$, $Z = 2.9$, $p = .01$).

Hence, performance in the control condition was slightly but significantly worse in the low-proficiency L2-English group than in either the high-proficiency L2-English group or the L1-Malay group.

Next, we analysed whether the three language groups differed in the target condition. Thus, we constructed a binomial generalised linear mixed-effects model predicting responses in the target condition ('true' or 'false') on the basis of language group, including random intercepts for participants and items. Again, we estimated the effect of the fixed factor using model comparison. Here, we found no significant effect of language group ($\chi^2(2) = 2.5$, $p = .29$).

Hence, the three language groups did not significantly differ in the probability with which they rejected underinformative sentences with 'some'. This conclusion is in line with earlier findings by Dupuy and colleagues (2019), among others.

Experiment 2: Computerised task

Participants

110 participants took part in this study. All participants were undergraduate students at Universiti Putra Malaysia. All but one of the participants were female. Their mean age was 22 (standard deviation: 2, range: 19–28).

As in Exp. 1, participants' English proficiency was determined on the basis of their band on the MUET. In our sample, 4 participants were in band 2, 56 in band 3, 45 in band 4, and 5 in band 5. Again, we grouped together participants in bands 2 and 3 into a low-proficiency group, and participants in bands 4 and 5 into a high-proficiency group.

The experiment comprised two blocks. One block tested Malay sentences; the other block tested English sentences. The order of the two blocks was counterbalanced: half of the participants started with the Malay sentences; the other half with the English sentences. For

comparison with Exp. 1, we focus exclusively on the data from the first block, before participants had seen sentences from the other language. However, all relevant analyses are confirmed when data from both blocks are analysed together.

Materials and procedure

The materials were the same as those used in Exp. 1. However, the procedure followed more closely the procedure used by Bott and Noveck (2004).

Participants were placed in front of a computer and were informed that they had to judge the truth value of categorical sentences by using keyboard buttons labeled ‘True’ and ‘False’. The button locations corresponded to the locations of the letters ‘c’ and ‘m’ on a QWERTY keyboard. The experiment was conducted using the E-prime software.

Each trial in the experiment consisted of a fixation cross and a sentence. The fixation cross remained on the screen for 500 milliseconds and was replaced by the sentence words that were consecutively flashed onto the screen, one word at a time. Each word remained on screen for 250 milliseconds, with a gap of 50 milliseconds between words. At the end of the sentence, participants were required to judge its truth value. They were instructed to respond as quickly as possible by pressing the corresponding button.

The experiment started with a practice session consisting of 18 trials. The practice session followed the same procedure as the main session but used sentences with different category and exemplar names. Participants were encouraged to ask questions during the practice session in order to work independently during the experimental session.

The truth-value judgement task consisted of two blocks. One block tested 54 L2-English sentences; the other block tested 54 L1-Malay sentences. For half of the participants, the L2-English sentences preceded the L1-Malay sentences; for the other half, it was the other way around. To avoid repetition effects, each block used a different list of test sentences. As

noted, we only report results from the first block; however, all relevant conclusions are confirmed when both blocks are analysed together.

The truth-value judgement task was followed by two tasks designed to measure participants' working memory capacity; namely, the Operation Span Task (OSpan, e.g., Shipstead et al., 2016) and the Symmetry Span Task (SymSpan, e.g., Kane et al., 2004; Unsworth et al., 2005). Here, we used the shortened versions of these two tasks, as published and validated by Foster and colleagues (2015).

The OSpan required participants to memorise sequences of letters from the Latin alphabet. The letters were separated by simple math problems. These math problems showed a mathematical equation that participants had to judge as being correct or incorrect (e.g., '(6 × 3) + 5 = 21'). Participants thus had to remember sequences of three, four, five, six, and seven letters. The order of the five trials was randomised. Participants responded using the keyboard.

The OSpan score equalled the number of letters that participants recalled in the correct order in which they appeared. The maximum score was 25 points. Higher OSpan scores indicated greater working memory capacity.

The SymSpan followed a similar procedure as the OSpan, but with some important differences. First, the to-be-remembered items were locations of red squares in a 4×4 grid. Second, the distractor task required participants to judge whether a displayed shape was symmetrical along its vertical axis. Participants had to memorise two sequences of two, three, four, and five locations. The order of the eight trials was randomised.

The SymSpan score equalled the number of red square locations that were correctly recalled. The maximum score was 28. Higher SymSpan scores indicated greater working memory capacity.

The OSpan always preceded the SymSpan. Both the OSpan and SymSpan started with a practice block to familiarise participants with the task. Based on their speed during the

practice trials, participants were automatically assigned an individualised response window corresponding to their mean response time on the practice trials plus 2.5 standard deviations.

All participants received a cash payment of $2.5 upon the completion of these tasks.

Data treatment

Twelve participants were removed from the analysis because they made mistakes in more than 20% of the control items.

Responses

Fig. 2 summarises the percentages of 'true' responses for each sentence type in each of the three language conditions (L1-Malay, high-proficiency L2-English, and low-proficiency L2-English). The percentages of 'true' responses in the target condition (i.e., sentence type T1) were 39%, 28%, and 55%, respectively.

<Insert Figure 2 about here>

First, we analysed whether the three language conditions differed in the control conditions (i.e., sentence types T2–T6). Thus, we constructed a binomial generalised linear mixed-effects model predicting responses (correct or incorrect) on the basis of condition (T2–T6), language group, and their interaction, including random intercepts for participants and items. The effects of the fixed factors were estimated using model comparison with more parsimonious models.

This analysis indicated that there were significant main effects of sentence type ($\chi^2(4) = 16.1$, $p = .003$) and language condition ($\chi^2(2) = 10.8$, $p = .005$). In addition, the interaction between these two factors was significant ($\chi^2(8) = 31.0$, $p < .001$).

Focusing on a model predicting responses on the basis of language condition, we pairwise compared the language conditions to determine which of them differed in terms of the proportion of correct responses. It was found that the proportion of correct responses was significantly lower in the low-proficiency L2-English group (88%) compared to the high-proficiency L2-English group (94%, β = -0.9, SE = 0.3, Z = -3.0, p = .007), and compared to the L1-Malay group (93%, β = 0.8, SE = 0.3, Z = 2.6, p = .02). The difference between L1-Malay group and the high-proficiency L2-English group was not significant (β = -0.1, SE = 0.3, $Z < 1$).

Hence, as in Exp. 1, performance in the control condition was significantly lower in the low-proficiency L2-English condition than in the high-proficiency L2-English and L1-Malay conditions. Visual inspection of Fig. 2 indicates that the source of this difference lies mainly in sentence types T2 and T5, possibly due to difficulties associated with verifying true or false superset sentences (McCloskey & Glucksberg, 1979).

Next, we analysed whether the three language conditions differed in the target condition. Thus, we constructed a binomial generalised linear mixed-effects model predicting responses (‘true’ or ‘false’) on the basis of language group, including random intercepts for participants and items. Again, we estimated the effect of the fixed factor using model comparison. This analysis indicated a significant effect of language group ($\chi^2(2) = 14.0$, $p < .001$).

We pairwise compared the language groups to determine which of them differed in terms of the proportion of ‘true’ responses in the target condition. It was found that the proportion of ‘true’ responses was significantly higher in the low-proficiency L2-English group than in both the high-proficiency L2-English group (β = 2.8, SE = 0.8, Z = 3.7, $p < .001$) and the L1-Malay group (β = -1.5, SE = 0.6, Z = -2.4, p = .04). The difference between the high-

proficiency L2-English group and the L1-Malay group was not significant (β = 1.3, SE = 0.6, Z = 2.0, p = .11).

Hence, unlike Exp. 1, we find in Exp. 2 that low-proficiency speakers of English are more likely to accept English underinformative sentences with 'some' in comparison to high-proficiency speakers, and in comparison to how likely native-speaking participants reject Malay underinformative sentences with 'some'.

Response times

Fig. 3 shows the mean response times in the target condition and in the control condition. The target condition is sentence type T1; the control condition comprises sentence types T2 (true sentences with 'some') and T3 (false sentences with 'some').

<Insert Figure 3 about here>

We analysed whether we confirmed Bott and Noveck's (2004) observation that 'false' responses took significantly longer than 'true' responses in the target condition, relative to the difference between these two responses in the control condition. Moreover, we wanted to see whether—if we replicated Bott and Noveck's finding—this effect was larger in the low-proficiency L2-English language condition compared to the L1-Malay and high-proficiency L2-English language conditions (Khorsheed, Rashid, et al., 2022).

Thus, we constructed a linear mixed-effects model predicting logarithmised response times on the basis of condition (target or control), response ('true' or 'false'), language condition (L1-Malay, high-proficiency L2-English, or low-proficiency L2-English), and their interactions, including random intercepts for participants and items. The effects of the fixed factors were estimated using model comparison with more parsimonious models. For this

analysis, degrees of freedom and corresponding *p*-values were estimated using the Satterthwaite procedure, as implemented in the 'lmerTest' package (Kuznetsova et al., 2013).

In line with Bott and Noveck's results, there was a significant interaction between condition and response ($\chi^2(1) = 29.0$, $p < .001$). However, this interaction did not interact with language condition ($\chi^2(6) = 8.3$, $p = .21$).

Hence, as Fig. 3 already suggested, the slowdown associated with rejecting underinformative sentences with 'some' was stable across the three language conditions, and, in particular, participants in the low-proficiency L2-English condition were not slowed down to a significantly greater degree than participants in the other two language conditions.

Working memory capacity

Since we did not have specific hypotheses about potential differences between the OSpan and SymSpan measures of working memory capacity, and since the two measures were significantly correlated ($r = .35$, $p < .001$), we calculated a general score for working memory capacity by taking the sum of the OSpan and SymSpan scores.

Fig. 4 provides a scatterplot showing working memory scores plotted against percentages of pragmatic responses (i.e., 'false' responses to underinformative sentences with 'some'). This figure suggests that greater working memory was associated with more pragmatic responses in L1-Malay, but not in either of the other language conditions.

<Insert Figure 4 about here>

To analyse whether working memory capacity significantly influenced participants responses in the target condition, we constructed a binomial generalised linear mixed-effects model predicting responses in the target condition on the basis of language condition (L1-

Malay, high-proficiency L2-English, low-proficiency L2-English), memory score, and their interaction, including random intercepts for participants and items. The effects of the fixed factors were estimated using model comparison with more parsimonious models.

There was a significant interaction between language condition and memory score ($\chi^2(2) = 7.1$, $p = .03$), but no overall main effect of memory score ($\chi^2(1) < 1$). To analyse the source of the significant interaction, we constructed, for each language condition separately, a binomial generalised linear mixed effects model predicting responses in the target condition on the basis of memory score. These analyses indicated a small but significant effect of memory score on responses for L1-Malay ($\beta = -0.1$, $SE = 0.1$, $Z = -2.3$, $p = .02$) but not for either of the other language conditions (high-proficiency L2-English: $\beta = 0.1$, $SE = 0.1$, $Z = 1.1$, $p = .27$, low-proficiency L2-English: $\beta = 0.1$, $SE = 0.1$, $Z < 1$). When interpreting these findings, it should be borne in mind that the significance of the first effect would disappear if the alpha level is corrected for multiple comparisons, e.g., using the Holm-Bonferroni method.

In any case, we did not confirm our hypothesis that working memory capacity would have a greater effect on the probability of pragmatic responses in the low-proficiency L2-English condition in comparison to the high-proficiency L2-English condition. Indeed, working memory capacity had a significant effect in neither of these conditions.

General discussion

Across many types of tasks and populations, experimental studies have shown that the computation of the scalar inference from ‘some’ to ‘not all’ is cognitively effortful (e.g., Bott & Noveck, 2004; De Neys & Schaeken, 2007). As a consequence, one might expect that people are less likely to compute scalar inferences in L2 than L1, since processing L2-input draws on more cognitive resources than L1-input. Surprisingly, many studies on scalar inferences in L2

fail to confirm this hypothesis, with some studies even showing that people are *more* likely to draw scalar inferences in L2 than L1 (e.g., Dupuy et al., 2019; Slabakova, 2010).

Following Mazzaggio and colleagues (2021) and Khorsheed, Rashid, and colleagues (2022), we proposed that the results from L2 studies have been confounded by the use of experimental tasks in which participants can process and evaluate sentences at their leisure. We hypothesised that, when participants cannot take their time to process and evaluate sentences, their difficulties with computing scalar inferences should come to the fore.

We tested this hypothesis on the basis of two experiments. In both experiments, participants had to evaluate sentences with ‘some’ or ‘all’. The target condition consisted of sentences such as (6), which were literally true but carried a scalar inference that was false.

(6) Some dogs are mammals.

Following the literature, we assumed that participants who computed the scalar inference would reject such underinformative sentences with ‘some’.

In Exp. 1, participants could take as much time as they wanted to read and evaluate the sentences. Here, we found no significant difference in the propensity with which participants computed scalar inferences in their L1 and L2, regardless of their L2 proficiency. By contrast, in Exp. 2, sentences were presented by flashing each word briefly on a computer screen, and participants were instructed to respond as quickly as possible because their response times were measured. In this experiment, we found that low-proficiency L2 speakers were significantly less likely to compute scalar inferences in their L2 compared to high-proficiency L2 speakers, and compared to the probability that L1 speakers computed scalar inferences in their L1.

However, unlike Mazzaggio and colleagues (2021), we found no significant differences between L1 speakers and high-proficiency L2 speakers in computing scalar inferences,

suggesting that difficulties in scalar inference comprehension can be overcome as L2 speakers have a stronger command of the target language. These results corroborate relevant reports in the literature (Destruel, 2022), as well as the idea that bilingualism enhances pragmatic understanding (Siegal et al., 2009, 2010), but they challenge the Interface Hypothesis, which states that L2 speakers experience difficulties with linguistic structures at the interfaces between structural or semantic levels (Sorace, 2011; Sorace & Filiaci, 2006). In contrast with the Interface Hypothesis, we did not observe a significant difference in the propensity with which L1 speakers and high-proficiency L2 speakers computed scalar inferences in their L1 and L2, even though scalar inferences are situated at the interface between semantics and pragmatics. Apparently, some L2 difficulties at the interface between semantics and pragmatics may disappear at higher levels of language attainment.

These findings confirm our hypothesis that the computation of scalar inferences in L2 is especially challenging for low-proficiency speakers, but that these difficulties may be masked if these speakers have the opportunity to take their time to process and evaluate the test sentences. Put differently, low-proficiency L2 speakers, due to their limited linguistic skills, may rely more on conscious and metalinguistic reasoning when given unlimited time to process and comprehend the target sentences, but they shift to more natural and intuitive reasoning when they are tasked to respond more quickly.

This observation aligns with Ellis (2009), who argues that pen-and-paper tasks may encourage L2 participants to engage in elaborate reasoning about sentences, and consequently provide more refined responses that only artificially match those of L1 speakers. By contrast, computerised tasks prompt participants to respond more intuitively. Therefore, the responses for the low-proficiency L2 speakers obtained from pen-and-paper experiments—such as our Exp. 1—were potentially a product of their explicit knowledge, whereas those obtained from

computerised experiments—including our Exp. 2—were reflective of implicit knowledge (see also Hopp, 2022, for relevant discussion).

There are at least two, mutually non-exclusive, explanations for our results. A first explanation is grounded in the relevance-theoretic idea that the computation of scalar inferences without contextual support draws on cognitive resources (e.g., Noveck & Sperber, 2007). According to this explanation, low-proficiency participants were less likely to derive scalar inferences in Exp. 2 because they had fewer cognitive resources at their disposal, since they had to keep in memory expressions from a language in which they had less proficiency when compared to the other two language groups.

There are at least three difficulties with this explanation. First, the explanation implies (or seems to imply) that scalar inferences should be consistently less frequent in L2 than in L1, given that processing L2 input is cognitively demanding (e.g., Clahsen & Felser, 2006; Green, 1986; 1998; Juffs, 2001; White & Juffs, 1998). This is not what has been observed in our study, or in most of the earlier studies on the topic (e.g., Dupuy et al., 2019; Feng & Cho, 2019). Second, the explanation suggests that any manipulations that introduce cognitive load should uniformly impact scalar inference rates, whereas we found that they impact low-proficiency language users to a greater extent than high-proficiency language users. Third, the proposed explanation is difficult to reconcile with the (admittedly few) studies that observed higher rates of scalar inferences in L2 (Slabakova, 2010; Snape & Hosoi, 2018).

According to a second explanation, there are two routes for computing scalar inferences. First, scalar inferences may be generated heuristically by recognising pragmatic regularities. However, this route requires substantial familiarity with language use, which is likely to be absent in low-proficiency speakers. Second, scalar inferences may be derived by reasoning about the speaker's intentions based on what they said and what they could have said. For example, the scalar inference of (6) may be derived by reasoning that, if the speaker

believed that all dogs are mammals, it would have been informative to say so. Hence, by saying (6), the speaker implicates they do not believe that all dogs are mammals.

To appreciate the heuristic route, consider the study by van Tiel and colleagues (2016). Van Tiel and colleagues observed substantial variability in the rates at which different scalar words trigger scalar inferences. Thus, whereas 'some' almost always implied 'not all', the inference from 'intelligent' to 'not brilliant', or from 'big' to 'not enormous', was found to be much less frequent. Although van Tiel and colleagues identified some structural factors that predicted the different rates of scalar inferences, a substantial part of the variability that they observed was left unexplained. As they recognise, it is likely that part of this unexplained variability is due to statistical regularities that are grounded in the way that language is used.

The idea that there are two routes to computing scalar inferences ties in with Grice's (1975, p. 50) observation that conversational implicatures can be "grasped intuitively" or "worked out" based on what the speaker said and the assumption that they are cooperative. Moreover, this idea might provide a synthesis between the relevance-theoretic idea that scalar inferences are particularised, i.e., context-dependent, and the defaultist idea that scalar inferences are automatically triggered by scalar expressions.

One way of teasing the relative contribution of these two explanations apart is by investigating effects of proficiency on the probability of deriving scalar inferences other than the inference from 'some' to 'not all'. If the explanation in terms of cognitive load is correct, any effects of proficiency should be relatively uniform across scalar expressions. By contrast, if the dual route explanation is correct, effects of proficiency may be more multifarious, i.e., less proficient language users may be, variously, more, equally, or less likely to derive scalar inferences than more proficient or native speakers, because they draw on different derivational pathways. We leave this line of inquiry for future research.

One of our reviewers pointed out that the switch from a pen-and-paper experiment to a computerised experiment also led to lower rates of scalar inferences in the L1 group (from 85% to 61%) and, to a lesser extent, in the high-proficiency L2 group (from 79% to 72%). Given that these two groups may employ the same two derivational pathways as the low-proficiency L2 group, it is not surprising that task demands also affected the propensity with which they derived scalar inferences. However, it is surprising that the effect of task demands is seemingly greater for the L1 group than for the high-proficiency L2 group.

We do not have a clear explanation for this finding. A potentially relevant observation is that only in the L1 group, the probability of computing scalar inferences was influenced by participants' working memory capacity. This asymmetry hints at a difference in the cognitive process that underlies the derivation of scalar inferences in L1 and L2. At the same time, further research is needed to confirm these findings, and to obtain a more fine-grained insight into the interaction between proficiency and the cognitive processing that underlies scalar inferencing.

Incidentally, the observation that increased task demands decrease the propensity with which L1 speakers compute scalar inferences ties in with the idea that the computation of scalar inferences is cognitively demanding (e.g., Bott & Noveck, 2004; De Neys & Schaeken, 2007).

In contrast with our hypotheses, and unlike Khorsheed, Rashid, and colleagues (2022), we did not observe significant effects of proficiency on the time needed for computing scalar inferences. As in Bott and Noveck (2004), we found that 'false' responses to sentences such as (6) took significantly longer than 'true' responses, and that such a difference was absent for control sentences that were unambiguously true or false. However, it was found that the magnitude of the delay was stable across the three language conditions.

Here, it should be noted that it has recently been questioned whether Bott and Noveck's finding really reflects the time needed for computing scalar inferences. For example, van Tiel, Pankratz, and Sun (2019) argue that this delay in response times instead reflects participants'

cognitive difficulties with the representation of negative information (e.g., the proposition that not all dogs are mammals).

Similarly, we did not confirm our hypothesis that low-proficiency speakers would be more sensitive to differences in working memory capacity than high-proficiency speakers when processing L2-input. Indeed, we did not observe significant effects of working memory capacity in either case. We did observe that working memory capacity had a significant positive relation to the probability of deriving scalar inferences, though this effect was small and became non-significant when the alpha level was corrected for multiple comparisons.

A potential limitation of our study is that, in Exp. 2, participants were tested in both their L1 and L2, whereas, in Exp. 1, participants were tested in only one of these languages. Recall that Dupuy and colleagues (2019) found that scalar inference rates were higher when participants were tested in both their L1 and L2 than when they were tested in only one language. In our experiments, we did not replicate this finding. Indeed, scalar inference rates were substantially higher in Exp. 1 than in Exp. 2 (81% vs. 60%). However, this difference may also be caused by the methodological differences between the two experiments. In any case, since our focus of inquiry was on the interaction between task demands and proficiency, rather than on effects of task demands per se, we believe that the presentational difference between Exps. 1 and 2 does not impact the conclusions that we draw from our study.

In summary, we have offered an explanation of the apparently anomalous finding that the computation of scalar inferences is independent of—or even inversely related to—language proficiency. We have shown that this finding relies on the use of experimental tasks that allow participants to reason about the test sentences in their L2 to compensate for their pragmatic difficulties. Thus, our study underlines the importance of taking seriously the affordances of experimental tasks in research on multilingualism.

Acknowledgements

We thank two anonymous reviewers for important comments on an earlier version of this article.

Ethical approval

All procedures received written approval from the ethical committee of Universiti Putra Malaysia (reference number: JKEUPM-2018-197). Informed consent was obtained from all individual participants included in the study.

Data availability

All data and analysis files can be accessed at OSF using the following link: https://osf.io/qchxn/?view_only=d6fc51bbb38f451795df8d4fd54d662e.

References

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

Antoniou, K., & Katsos, N. (2017). The effect of childhood multilingualism and bilectalism on implicature understanding. *Applied Psycholinguistics*, *38*, 787–833.

Antoniou, K., Veenstra, A., Katsos, N., & Kissine, M. (2018). How does childhood bilingualism and bi-dialectalism affect the interpretation and processing of pragmatic meanings? *Bilingualism: Language and Cognition*, *23*, 186–203.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

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

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

Carston, R. (2004). Truth-conditional content and conversational implicature. In C. Bianchi (Ed.), *The semantics/pragmatics distinction* (pp. 65–100). CSLI Publications.

Chemla, E., & Bott, L. (2014). Processing inferences at the semantics/pragmatics frontier: Disjunctions and free choice. *Cognition*, *130*, 380–396.

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics*, *27*, 3–42.

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

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

Destruel, E. (2022). Processing pragmatic inferences in L2 French speakers. *Second Language Research*, *39*, 969–995.

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

Dupuy, L., Stateva, P., Andreetta, S., Cheylus, A., Dèprez, V., van der Henst, J.-B., ... Reboul, A. (2019). Pragmatic abilities in bilinguals: The case of scalar implicatures. *Linguistic Approaches to Bilingualism*, *9*, 314–340.

Ellis, R. (2009). Measuring implicit and explicit knowledge of a second language. In *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 31–64). Multilingual Matters.

Feng, S., & Cho, J. (2019). Asymmetries between direct and indirect scalar implicatures in second language acquisition. *Frontiers in Psychology*, *10*, 1–17.

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory and Cognition*, *43*, 226–236.

Gazdar, G. (1979). Pragmatics: Implicature, presupposition, and logical form. Academic Press.

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

Geurts, B., & Rubio-Fernández, P. (2015). Pragmatics and processing. *Ratio*, *28*, 446–469.

Green, D. W. (1986). Control, activation and resource: A framework and a model for the control of speech in bilinguals. *Brain and Language*, *27*, 210–223.

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, *1*, 67–81.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, volume 3: Speech acts* (pp. 41–58). 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*, 42–55.

Holtgraves, T., Kwon, G., & Zelaya, T. M. (2019). Psycholinguistic approaches to L2 pragmatics research. In N. Taguchi (Ed.), *The Routledge handbook of second language acquisition and pragmatics* (pp. 272–284). Routledge.

Hopp, H. (2022). Second language sentence processing. *Annual Review of Linguistics*, *8*, 235–256.

Horn, L. R. (1972). *On the semantic properties of logical operators in English* (Unpublished doctoral dissertation). University of California, Los Angeles.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*, 346–363.

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

Janssens, L., Fabry, I., & Schaeken, W. (2014). 'Some' effects of age, task, task content and working memory on scalar implicature processing. *Psychologica Belgica*, *54*, 374–388.

Juffs, A. (2001). Psycholinguistically-oriented second language research. *Annual Review of Applied Linguistics*, *21*, 207–223.

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*, 189–217.

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

Khorsheed, A., Price, J., & van Tiel, B. (2022). Sources of cognitive cost in scalar implicature processing: A review. *Frontiers in Communication*, *7*, 990044.

Khorsheed, A., Rashid, S. M., Nimehchisalem, V., Imm, L. G., Price, J., & Ronderos, C. R. (2022). What second-language speakers can tell us about pragmatic processing. *PLOS ONE*, *17*, e0263724.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2013). lmerTest: tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package) [R package]. Retrieved from http://cran.r-project.org/package=lmerTefittst

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

Lin, Y. (2016). Processing of scalar inferences by Mandarin learners of English: An online measure. *PLOS ONE*, *11*, 1–27.

Marty, P., Chemla, E., & Spector, B. (2013). Interpreting numerals and scalar items under memory load. *Lingua*, *133*, 152–163.

Mazzaggio, G., Panizza, D., & Surian, L. (2021). On the interpretation of scalar implicatures in first and second language. *Journal of Pragmatics*, *171*, 62–75.

McCloskey, M., & Glucksberg, S. (1979). Decision processes in verifying category membership statements: Implications for models of semantic memory. *Cognitive Psychology*, *11*, 1–37.

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

Noveck, I. A., Fogel, M., Van Voorhees, K., & Turco, G. (2022). When eleven does not equal 11: Investigating exactness at a number's lower bound. *PLOS ONE*, *17*, e0266920.

R Core Team. (2023). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.r-project.org/

Romoli, J., & Schwarz, F. (2015). An experimental comparison between presuppositions and indirect scalar implicatures. In F. Schwarz (Ed.), *Experimental perspectives on presuppositions* (pp. 215–240). Springer.

Ronai, E., & Xiang, M. (2021). Pragmatic inferences are QUD-sensitive: An experimental study. *Journal of Linguistics*, *57*, 841–870.

Schaeken, W., Van de Weyer, L., De Hert, M., & Wampers, M. (2021). The role of working memory in the processing of scalar implicatures of patients with schizophrenia spectrum and other psychotic disorders. *Frontiers in Psychology*, *12*, 635724.

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.

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2016). Working memory capacity and fluid intelligence: Maintenance and disengagement. *Psychological Science*, *11*, 771–799.

Siegal, M., Iozzi, L., & Surian, L. (2009). Bilingualism and conversational understanding in young children. *Cognition*, *110*, 115–122.

Siegal, M., Surian, L., Matsuo, A., Geraci, A., Iozzi, L., Okumura, Y., & Itakura, S. (2010). Bilingualism accentuates children's conversational understanding. *PLOS ONE*, *5*, e9004.

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*, 2444–2462.

Snape, N., & Hosoi, H. (2018). Acquisition of scalar implicatures: Evidence from adult Japanese L2 learners of English. *Linguistic Approaches to Bilingualism*, *8*, 163–192.

Sorace, A. (2011). Pinning down the concept of "interface" in bilingualism. *Linguistic Approaches to Bilingualism*, *1*, 1–33.

Sorace, A., & Filiaci, F. (2006). Anaphora resolution in near-native speakers of Italian. *Second Language Research*, *22*, 339–368.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Blackwell.

Sperber, D., & Wilson, D. (2006). Relevance theory. In L. R. Horn & G. Ward (Eds.), *The handbook of pragmatics* (pp. 607–632). Blackwell.

Spychalska, M., Kontinen, J., & Werning, M. (2016). Investigating scalar implicatures in a truth-value judgement task: Evidence from event-related brain potentials. *Language, Cognition and Neuroscience*, *31*, 817–840.

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

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*, 498–505.

van Tiel, B., & Pankratz, E. (2021). Adjectival polarity and the processing of scalar inferences. *Glossa*, *6*, 32.

van Tiel, B., Pankratz, E., & Sun, C. (2019). Scales and scalarity: Processing scalar inferences. *Journal of Memory and Language*, *105*, 427–441.

van Tiel, B., & Schaeken, W. (2017). Processing conversational implicatures: Alternatives and counterfactual reasoning. *Cognitive Science*, *41*, 1–36.

van Tiel, B., van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016). Scalar diversity. *Journal of Semantics*, *33*, 137–175.

White, L., & Juffs, A. (1998). Constraints on wh-movement in two different contexts of non-native language acquisition: Competence and processing. In S. Flynn, G. Martohardjono, & W. O'Neill (Eds.), *The generative study of second language acquisition* (pp. 111–130). Erlbaum.

Yang, X., Minai, U., & Fiorentino, R. (2018). Context-sensitivity and individual differences in the derivation of scalar implicature. *Frontiers in Psychology*, *9*, 1720.

Label	Example sentence	Correct response
T1	Some parrots are birds	?
T2	Some birds are parrots	T
T3	Some parrots are fish	F
T4	All parrots are birds	T
T5	All birds are parrots	F
T6	All parrots are fish	F

Table 1: Example test sentences used in Exps. 1 and 2.

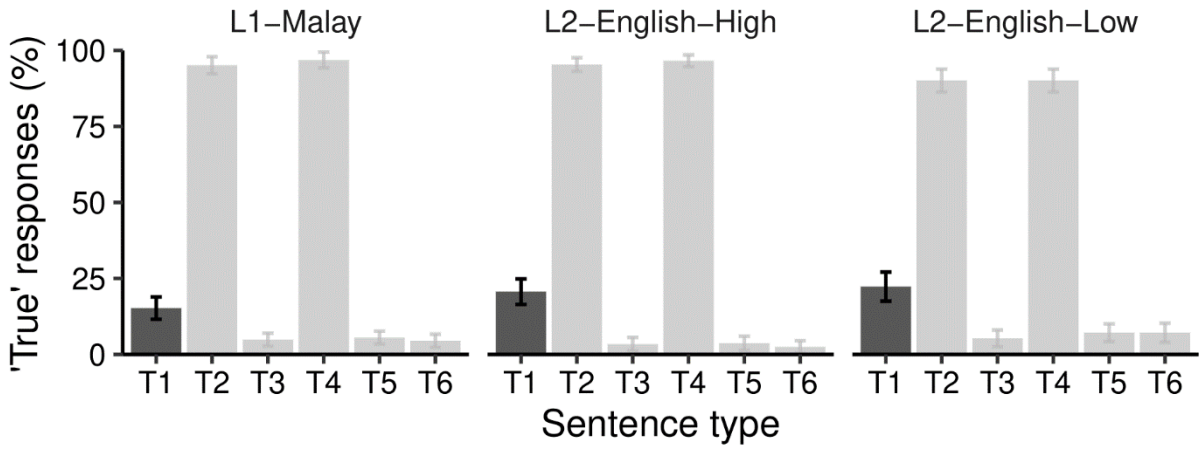

Figure 1: Percentage of 'true' responses for each sentence type in each language group (Exp. 1). T1 is the target condition consisting of underinformative sentences with 'some'. Error bars represent 95% confidence intervals.

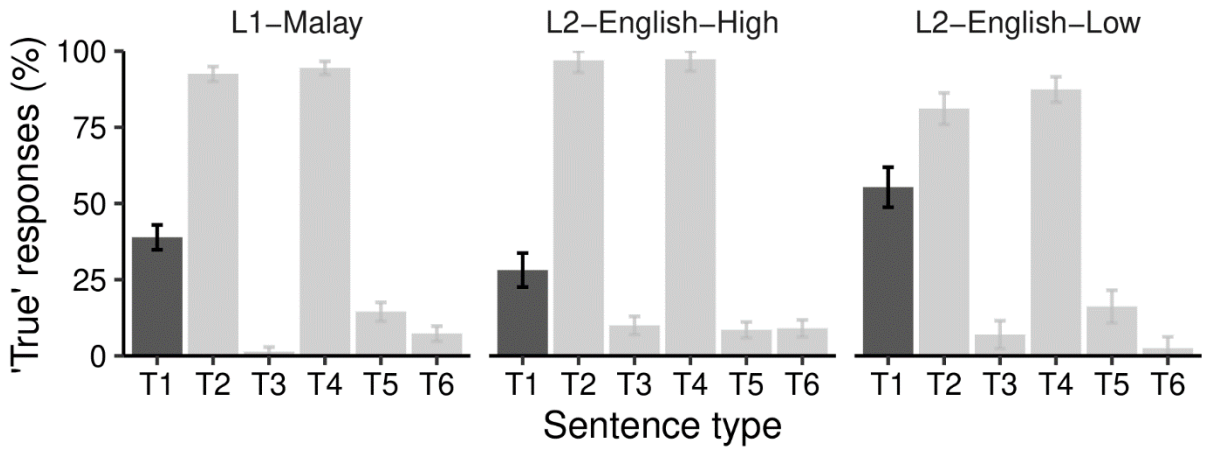

Figure 2: Percentage of 'true' responses for each sentence type in each language condition (Exp. 2). T1 is the target condition consisting of underinformative sentences with 'some'. Error bars represent 95% confidence intervals.

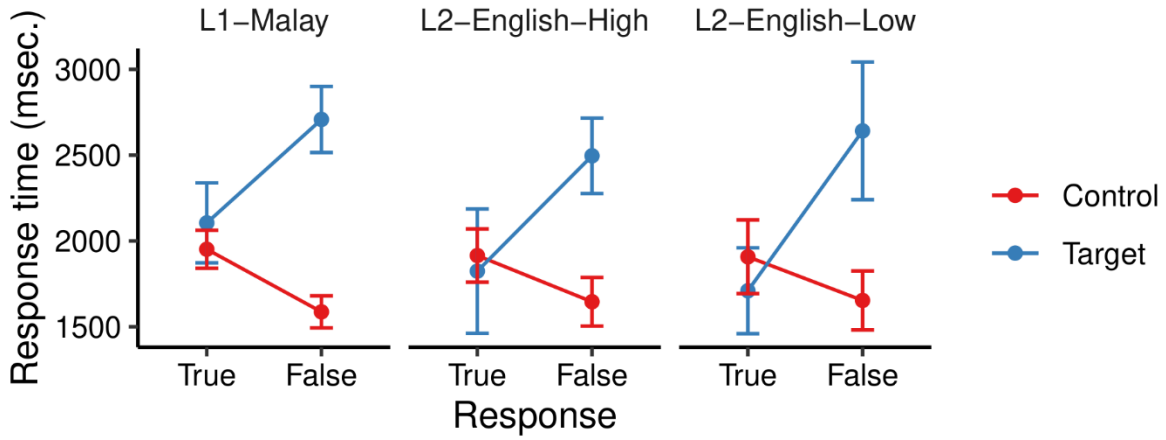

Figure 3. Mean response times in the target condition (T1) and control condition (T2 and T3) in each language group. Errors bars represent 95% confidence intervals.

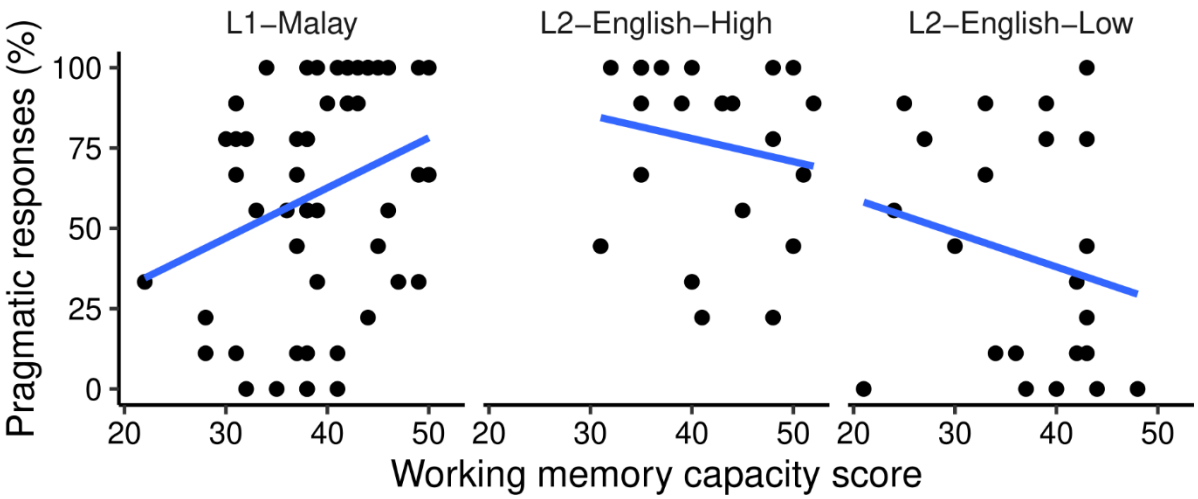

Figure 4. Scatterplot showing working memory capacity scores against pragmatic (i.e, 'false') responses in the target condition for each language condition in Exp. 2.