Applied Psycholinguistics (2024), **45**, 717–744
doi:10.1017/S0142716424000274

CAMBRIDGE UNIVERSITY PRESS

ORIGINAL ARTICLE

Pragmatic competence and pragmatic tolerance in foreign language acquisition—revisiting the case of scalar implicatures

Johannes Schulz and Elizabeth Wonnacott

Department of Education, University of Oxford, Oxford, Oxfordshire, UK
Corresponding author: Johannes Schulz; Email: johannes.schulz@education.ox.ac.uk

(Received 8 August 2023; revised 26 April 2024; accepted 9 June 2024; first published online 23 September 2024)

Abstract

Previous L2 studies used binary Truth-Value-Judgment (TVJ) tasks to investigate L1–L2 differences in scalar implicature derivation (*some X* implicates *some but not all X*). They examined participants' judgments of sentences with weak scalar expressions ("Timothy ate some of the pretzels") when stronger ones are true ("Timothy ate all of the pretzels"). Some studies indicate adult L2 learners are less likely than L1 users to accept such statements while others found the opposite, concluding that implicature derivation is "costly" for L2 learners, rendering them less pragmatically competent than L1 users. Importantly, related L1 research suggests that TVJ tasks only capture sensitivity to under-informativeness. This sensitivity might be completely overridden by metalinguistic attitudes in binary tasks, whereas graded tasks reveal nuanced judgment patterns. Exploring L2 response behaviors, we tested English L1 speakers and competent German L2 English learners using binary and graded tasks. In both tasks, we found evidence of pragmatic responding with no evidence of differences between groups. Bayes factor analyses of the graded data favored H0 over the hypotheses that L2 learners provide fewer or more rejections to under-informative input than L1 learners. We explore implications for L2 learners' pragmatic abilities, differences with previous studies, and the role of TVJ tasks in under-informative contexts.

Keywords: Foreign language learning; pragmatic competence; pragmatic tolerance; scalar implicatures; sensitivity to under-informativeness; Truth-Value-Judgment (TVJ)

Introduction

Understanding language involves making inferences. People frequently convey more meaning with what they say than is conveyed by the meaning of the words:

(1) a. A: Did Sophie eat all of the cupcakes?
b. B: Sophie ate some of them.
c. *Implicature: Sophie did not eat all of the cupcakes.*

© The Author(s), 2024. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

In (1), B is not as informative as required by the communicative purpose. According to Grice (1975) and other reformulations of his theory (Geurts, 2010; Levinson, 2000), this leads A to derive the implicature (1c) in a process which is two-staged: first, A determines whether B could have been more informative (*I ate all of the cupcakes*). Given the question (1a), the additional information would be relevant. This *sensitivity to informativeness* is the necessary first step of implicature derivation (Grice, 1975; Geurts, 2010; Katsos, 2009). Second, A assumes B to be cooperative and maximally informative; hence, B would have used the more informative term *all* if B believed it were true. A also assumes B to know about the truth of the stronger alternative (*Epistemic Step*, cf. Sauerland, 2004). Since B used the less informative term (*some*), A infers the negation of the stronger term (*all*) on the informativeness scale <*some-many-all*> (Horn, 1972), deriving a *scalar implicature*.

While various theories of the processing mechanisms underlying scalar implicatures are available (Sauerland, 2012), L2 research in this area has exclusively employed the neo-Gricean *default* (or *lexical*) and post-Gricean *non-default* models to account for the cognitive aspects of this inferencing. The default model (Horn, 2004; Horn et al., 2012; Levinson, 2000) proposes an automatic and immediate pragmatic enrichment (*some but not all*) of scalar terms irrespective of communicative context. Thus, pragmatic interpretations require less processing effort than logical interpretations because the former need to be canceled first before the latter can be derived (Dupuy et al., 2019). In contrast, the non-default model (Sperber & Wilson, 1986; Carston, 1998; Noveck & Sperber, 2007) proposes no default pragmatic enrichment of scalar terms. Instead, recipients process the entire sentence and derive the logical meaning. Subsequently, context conditions warranting an implicature must be satisfied to derive the pragmatically enriched meaning. Therefore, pragmatic interpretations add processing effort (Dupuy et al., 2019).

In experiments investigating the link between existentially quantified sentences like (1b) and the corresponding scalar implicature in (1c), conversations like (1a–b) would usually be situated in scenarios where it is obvious that "Sophie has eaten all of the cupcakes," rendering her response in (1b) *under-informative* yet logically true. In much L2 research (cf. L1 research: Noveck, 2001; Noveck & Posada, 2003; Bott & Noveck, 2004; Feeney et al., 2004; de Neys & Schaeken, 2007; Pijnacker et al., 2009), under-informative utterances have been used as targets in binary Truth-Value-Judgment (TVJ) tasks to elicit responses from participants (*yes/no*; *agree/disagree*). Usually, rejections of under-informative target sentences such as (1b) or (2) "Some elephants have trunks" (i.e., drawing on encyclopedic knowledge) have been assumed to indicate that participants derived the scalar implicature (i.e., *some, but not all* interpretation), warranting a rejection. Conversely, failure to derive the scalar implicature would be interpreted to indicate that the participant employed the logical alternative (i.e., *some and possibly all*), warranting acceptance. Experimental predictions assumed that L2 learners, even at advanced proficiency, have less language processing resources available than native speakers (Clahsen & Felser, 2006). Thus, under the default view, L2 speakers should make more pragmatic interpretations than L1 controls, and *vice versa* under the non-default view. Results have been inconsistent, with some reporting more implicature derivation among L2 learners compared to L1 speakers (Lin, 2016; Slabakova, 2010; Snape & Hosoi, 2018), supporting the *default model*. Others report no differences between L2

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

learners and native speakers (Dupuy et al., 2019), and some report that L2 learners derive fewer implicatures than native speakers (Mazzaggio et al., 2021; Khorsheed et al., 2022), supporting the *non-default model*.

Truth-Value-Judgment (TVJ) tasks and implicature derivation

Katsos and Bishop (2011) were among the first to scrutinize the use of binary TVJs when measuring implicature derivation in L1 research (L1 research discussed in 1.2). They suggest that rejecting *some X are Y* in a context where clearly *all X are Y* only requires participants to realize "that a more informative statement could have been made" (ib.: 69). This sensitivity to under-informativeness is considered sufficient to warrant rejection, without the need to derive the negation of the stronger alternative, constituting the implicature. Corresponding to Katsos and Bishop's suggestion (for which their 2011 report supplies no empirical evidence), Kissine and de Brabanter (2023: 3) formulate the *No-Implicature Hypothesis* stating that "judging a sentence of the form *Some X are Y* to be false in a situation where it is obvious that *all X are Y* does not entail the explicit representation of the reinforced reading *Some, but not all X are Y*." Importantly, they argue that scalar implicature research employing the default and non-default theories (among other theories) of implicature derivation all endorse the contrasting *Explicit Implicature Hypothesis*, stating that "judging a sentence of the form *Some X are Y* to be false in a situation where it is obvious that *all X are Y* entails explicitly representing the reinforced reading *Some, but not all X are Y*" (ib.: 2). Kissine and de Brabanter (2023) tested these hypotheses[1] in a series of experiments where participants judged under-informative sentences in binary TVJ tasks like those employed in previous L1 and L2 research. Each TVJ was immediately followed by a forced-choice between two paraphrases (logical and pragmatic reading) of the previously encountered under-informative sentences. In line with the *No-Implicature Hypothesis*, the authors found that participants overwhelmingly chose the logical paraphrase of under-informative sentences irrespective of their initial judgment of them. Their evidence strongly suggests that rejections in binary TVJ tasks "occur without the corresponding implicature being derived" (p.11) and that rejections are triggered by participants' sensitivity of the target sentence's oddness in their respective contexts.[2] The authors further highlight that those judgments of pragmatic violations are impacted by metalinguistic attitudes, which will become important later.

In contrast to Kissine and de Brabanter's (2023) findings, the L2 scalar implicature studies to date employing TVJ tasks have assumed that rejections indicate implicature derivation. This raises the question whether the adjudications between different implicature theories in previous L2 research and any claims made about participants' pragmatic competence regarding their implicature derivation ability may be unfounded since the studies might not have measured implicature derivation after all. Regarding pragmatic competence, all that binary TVJ tasks may have measured is participants' sensitivity to under-informativeness (i.e., the pragmatic violation). A similar argument has been put forward in L1 child research, discussed below.

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

Child–adult differences in binary TVJ tasks

Many L1 studies investigating children's implicature derivation development have employed binary tasks. Noveck (2001) initiated this line of inquiry when he had 8-to-10-year-olds make binary judgments of under-informative sentences (e.g., *some giraffes have long necks*) in a sentence verification task. Unlike adults, children did not reject under-informative input, albeit showing adult-like performance on logically false sentences. Subsequent studies introduced various paradigm manipulations to influence children's rejection and acceptance rates of under-informative input, including implicature relevance (Feeney et al., 2004), other under-informative sentence types (Papafragou & Tantalou, 2004; Barner et al., 2011), or explicit instruction and training (Papafragou & Musolino, 2003; Guasti et al., 2005). Although we acknowledge that other studies employed fundamentally different experimental paradigms (e.g., Pouscoulous et al., 2007), the critical point about the aforementioned studies using binary TVJ tasks is that "these studies base their conclusions on the assumption that the children who accept under-informative utterances altogether lack (some or all of the) competence that is needed to generate implicatures" (Veenstra et al., 2018: 299). This resembles the assumptions made in L2 research regarding the pragmatic abilities of L2 users.

Pragmatic tolerance

Scrutinizing the cause of children's acceptance of pragmatic violations, Veenstra and Katsos (2018; cf. Davies & Katsos, 2010; Katsos & Smith, 2010; Veenstra et al., 2018; Katsos & Bishop, 2008; 2011) suggest that children's apparent inability to interpret under-informative sentences pragmatically could stem from researchers' overinterpretation of binary data. They argue that agreement with under-informative statements ("Some elephants have trunks") does not *necessitate* implicature derivation failure. Instead, children may be sensitive to the under-informativeness, while still being willing to *tolerate*—and thus accept—the less felicitous interpretation since they might not consider the violation grave enough to warrant rejection (Schmitt & Miller, 2010).

These claims regarding pragmatic *tolerance* stem from Katsos and Bishop (2011) who tested adults and children using binary and ternary judgments. Ternary judgments offered an additional "intermediate" decision possibility. Participants observed scenes such as a giraffe eating all of the pears from a tree but not the apples (the tree bore both fruits). After exposure, experimenters asked an animated figure *What did the giraffe eat?* In under-informative trials, it would answer *The giraffe ate some of the pears*. In experiment 1, participants rewarded the figure's answers by evaluating them with *that's right* or *that's wrong* (binary response). In experiment 2 (same input), participants could choose between three different sized strawberries (small, medium, large; ternary judgment). Middle-sized strawberries allowed for an intermediate endorsement level. Experiment 1 confirmed previous findings (e.g., Noveck, 2001): children provided significantly more logical responses than adults in under-informative scenarios, demonstrating "lower" pragmatic competence. However, in experiment 2 with ternary response options, no evidence was found for adult–child differences and *both* groups demonstrated sensitivity to under-informativeness, favoring intermediate judgments in under-informative

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

contexts. The authors argue that intermediate responses enabled participants to demonstrate they noticed the pragmatic violations without enforcing categorical judgments. This appeared to remove adult–child differences, suggesting that previous differences were due to differences in pragmatic tolerance rather than the ability to detect under-informativeness.[3] In a similar study using quinary judgments (five-point Likert scales), Katsos and Smith (2010) confirmed that children demonstrate sensitivity to under-informativeness by choosing intermediate options in under-informative trials. Those findings brought about the *Pragmatic Tolerance Hypothesis* suggesting that children are sensitive to under-informativeness, but they differ with adults in their pragmatic tolerance, decreasing their tendency to reject such input.

Before reviewing previous L2 research, let us recap three issues highlighted in L1 research that are important to L2 research: first, TVJ tasks might not capture implicature derivation but merely sensitivity to under-informativeness. Mapped onto L2 literature, this means that data collected with TVJ tasks might not be able to adjudicate between competing implicature derivation theories. Moreover, although sensitivity to under-informativeness is potentially costly since it involves a consideration of the potential for a more informative statement (Katsos, 2009), it is presumably less costly than "fully fledged" implicatures and unlikely to exceed the cognitive capacity of L2 users, despite the L2 burden.[4] Thus, it would be reasonable to expect no differences to emerge between L1 and L2 users at the level of sensitivity to under-informativeness. Second, binary scales might rather shed light on pragmatic tolerance while potentially hiding participants' true ability to be sensitive to under-informativeness ("implicature derivation" is not captured in any case; Kissine & de Brabanter, 2023; Katsos & Bishop, 2011). The proposed developmental differences regarding pragmatic tolerance between adults and children seem plausible, yet, turning to L2 research, they do not map onto L1 and L2 users. If binary TVJ tasks outcomes are heavily impacted by aspects of pragmatic tolerance, we would not expect L1–L2 differences (with adult speakers) caused by cognitive development in this respect. This raises the question which other factors unrelated to cognitive and linguistic ability caused L1–L2 differences in response behavior in binary tasks in the past. Third, L1 research suggests that graded tasks show a more nuanced picture of the balance between pragmatic and logical responding. Thus, they might present some interesting insights when investigating L1 and L2 users' response patterns.

Previous L2 studies

We will now present an outline of previous L2 research. Recall that earlier in this paper, we introduced two theoretical accounts of implicature derivation (*default* and *non-default*) because those are important to trace previous L2 studies' theoretical motivations and their respective conclusions given their data. Based on our discussion of the L1 literature, we acknowledge that previous L2 studies employing TVJ tasks might not have shed light on implicature derivation.

Slabakova (2010) compared judgments of under-informative sentences by English L1 speakers, Korean L2 English learners (intermediate and advanced), and Korean L1 speakers, employing four different sentence types. Experiment 1 included

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

sentences without visual context drawing on encyclopedic knowledge ("Some elephants have trunks"). Experiment 2 included visual context (picture-stories featuring a woman and a girl). In both experiments, all language groups consistently provided "agree" responses to optimally true *all* and felicitous *some* sentences and "disagree" responses to optimally false *all* sentences. Under-informative sentences were critical trials. Participants were more likely to accept these in the encyclopedic than the visual context condition. Critically, in both experiments, L2 learners gave significantly more "disagree" responses to under-informative sentences (i.e., pragmatic responses) than L1. Lin (2016) used similar methods to test Mandarin L2 English learners' (CEFR[5]: B1-B2) judgments of under-informative *some* sentences in binary judgments and also reports that L2 learners gave significantly more pragmatic responses than L1 controls.

Snape and Hosoi (2018) compared Japanese L2 English learners (B2) with L1 controls and found that L2 learners tended to respond more pragmatically than controls, which they argue supports the default model. Since the results were not statistically significant, this should be treated cautiously.

Investigating this L2 "pragmatic bias," Dupuy et al. (2019) asked French monolinguals, French L2 English learners (B2), and French L2 Spanish learners (B2) to give *Yes/No* responses to under-informative statements (*The boy has hidden some cars*) provided visual context. L2 learners were tested in both their L1 and L2. Both L2 groups gave more pragmatic answers than monolinguals, but *only* when tested in both languages within the same session. Critically, when this occurred, they demonstrated biases in both their L1 (French) and L2 (English/Spanish). The authors argue that this is *not* consistent with an account attributing pragmatic biases to L2 speakers' processing disadvantages and that the findings could be attributed to an intra-experimental language switch which temporarily increased participants' metalinguistic awareness, including their susceptibility to pragmatic interpretations in their L1. Furthermore, they argue that L2 learning might generally increase speakers' metalinguistic awareness and compensate for the lack of proficiency compared to native speakers, resulting in equally "pragmatic" responses in native and non-native languages.

Contrary to earlier research, a recent study reports *less* pragmatic responding among L2 users. Mazzaggio et al. (2021) tested Italian L1 speakers and Italian L2 English and Spanish learners employing binary judgments. Contrary to previous studies, input was aural only, and time limits were introduced. The increased processing difficulty should potentially amplify the effect that processing difficulties lead to more pragmatic responses in L2 learners. However, they provided *fewer* pragmatic answers compared to L1 controls. The authors argue that their findings differ from Slabakova's (2010) because Slabakova's L2 participants were immersion students which might have improved their pragmatic abilities (Bouton, 1992). Yet, it remains unclear why immersion students "outscored" L1 controls and derived *more* pragmatic interpretations compared to Mazzaggio et al. (2021), rather than finding no differences.

Previous L2 findings are inconclusive. Yet, researchers made assertive claims about L1 and L2 speakers' implicature derivation processes and pragmatic abilities. While the individual reports provide reasons for the differences between them, those reasons assume that the binary tasks provided information about some aspect of the

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

participants' cognitive abilities. Considering that TVJ tasks may not capture implicature derivation but rather only sensitivity to under-informativeness, we argue that there is no reason to expect that L2 users are insensitive to pragmatic violations (as was indicated by "agree" responses in Mazzaggio et al., 2021), and therefore there is no reason to expect L1–L2 differences. Instead, we suggest that one possibility to account for differences in response behavior might be differing metalinguistic attitudes. Veenstra and Katsos (2018: 271) highlight the "fundamental challenge of any [TVJ], namely that it involves a meta-linguistic judgment." This notion that metalinguistic attitudes guide decision-making processes is reminiscent of related L2 experimental pragmatics research where it has been repeatedly argued that the attitudes impacting such decisions can be shaped by cross-linguistic and contextual factors. Those may include the cross-linguistic differences in the quantification of "some" (Stateva et al., 2019), the employed quantifier structures (Degen & Tanenhaus, 2013; Grodner et al., 2010), the perceived speaker likeability (Sikos et al., 2019), the discourse context (Yang et al., 2018; Dupuy et al., 2016), the type of judgment question asked (i.e., truth- vs. felicity-judgments; Kissine & de Brabanter, 2023), the prosody in aural input (Chen et al., 2018; Bill et al., 2018), or the perceived honesty traits of speakers (Feeney & Bonnefon, 2012). For example, Feeney and Bonnefon (2012) found effects of politeness contexts and individuals' honesty traits on the interpretation of scalar expressions. They demonstrated that in face-threatening contexts their participants gave fewer pragmatic answers compared to non-face-threatening contexts. They also provided evidence that participants gave more pragmatic interpretations the higher they rated their self-perceived honesty (regardless of context). Studies like this demonstrate that there are impactful factors other than cognitive that might "push" participants to accept or reject pragmatic violations.

Considering that binary responses might be heavily influenced by metalinguistic attitudes, we propose the possibility that L1–L2 differences observed in previous L2 studies might not have to do with pragmatic abilities but rather reflect a mix of participants' metalinguistic attitudes shaped by their linguistic backgrounds and contextual factors.

The current study[6]

We replicate previous L2 studies using a binary TVJ task and supplement the research with a graded task. Unlike previous L2 research, we reframe the inquiry to investigating sensitivity to pragmatic violations (i.e., under-informativeness) and acceptance thereof instead of implicature derivation abilities. Since previous L2 studies made assertive claims about L2 users' pragmatic abilities, we first and foremostly aim to provide evidence for the (perhaps unsurprising) fact that both L1 and L2 users are sensitive to under-informativeness and that there are no between-group differences at this fundamental level of pragmatic ability. In this regard, the graded task does not necessarily yield additional information regarding sensitivity to under-informativeness since every rejection in binary tasks, no matter how strong the impact of metalinguistic attitudes, already requires sensitivity to the pragmatic violation in the first place.[7] Nonetheless, the graded task provides additional, more nuanced insights into the balancing act between sensitivity to pragmatic violations

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

on the one hand and acceptance of such violations on the other hand. It sheds light on the trade-offs that participants from the different language groups are willing to make between the two poles. Such details are masked in the binary data; thus, the graded data ideally fathom and expand the insights gained from the binary data.[8] Taken together, both the binary and the graded task contribute to our investigation of participants' judgments of pragmatic violations and might inform future studies.

In both tasks, we expect both groups to demonstrate sensitivity to under-informativeness and we expect no between-group differences in this respect. In addition, since we keep contextual factors unmanipulated, we also have no reason to expect that proficient L2 users and L1 users differ in their judgment of such pragmatic violations. As a result, response behaviors should not differ between groups in both the binary and the graded task. To ensure consistency with previous studies, we test competent L2 learners (>B2), using newer, more appropriate statistical tools—Bayes factor (BF) statistics—to test evidence for H0. Finding such evidence would impact our perspective on L1–L2 differences with regard to their sensitivity to under-informativeness and their tolerance thereof.

Regarding our experimental hypotheses, recall that previous L2 studies employing binary tasks, although having framed their work against the background of implicature derivation, report findings in both directions, that is, L2 users provide more rejections (Slabakova, 2010) or L2 users provide fewer rejections (Mazzaggio et al., 2021). Therefore, we test the binary data in both directions separately (i.e., using two one-tailed tests; Ziori & Dienes, 2015). Note that although we believe it likely that the null hypothesis will be supported, testing a BF requires us to test a model of H1. The H1s that we will test are that L2 learners give fewer pragmatic responses than L1 learners and that L2 learners give more pragmatic responses.

There is an absence of prior L2 research using graded scales, yet as rightly argued by reviewers, the graded task—albeit more nuancedly so—taps into the same decision-making process as the binary task. Therefore, we test the graded data in both directions separately as well (i.e., using two one-tailed tests). Again, we will test whether L2 learners give fewer pragmatic responses than L1 learners and whether L2 learners give more pragmatic responses.[9]

All analyses were performed using R Statistical Software (v4.1.2; R Core Team 2021).

Design

The between-participant factor was "language group": native English monolinguals and German L2 English learners. Participants rated sentence-picture correspondence on either a binary (*Disagree* and *Agree*) or a five-point scale (*Disagree*, *Somewhat Disagree*, *Neither*, *Somewhat Agree*, and *Agree*). There were four sentence types: optimally true *all*, optimally false *all*, felicitous *some*, and under-informative *some*. Under-informative trials are the targets. In binary trials, high agreement suggests higher acceptance of pragmatic violations, while disagreement implies lower acceptance. In quinary tasks, strong agreement tendencies indicate higher acceptance, while strong disagreement tendencies suggest lower acceptance. Any response other than full agreement indicates sensitivity to under-informativeness in both tasks. While we could test our hypotheses with analyses conducted exclusively

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

on responses to under-informative items, high/low scores on this measure may also reflect broader preferences to agree/disagree more generally in the experimental context at the upper/lower ends of the Likert scale, rather than pragmatic sensitivity.[10] To control for this, we conducted analyses which compared responses to under-informative items to those to felicitous items (where high agreement levels are expected). Data from optimally true/false *all* trials were not analyzed. Instead, high performance on those trials served as baseline inclusion criterion in the final dataset (as in Dupuy et al., 2019).

Participants

Participants were recruited via Prolific and randomly allocated to either the experiment featuring the binary or the graded task. For the binary condition, the final sample comprised 55 English monolinguals and 54 German L2 English learners. One additional L2 participant was excluded for scoring below 75% correct in optimally true/false *all* trials. The remaining L2 participants demonstrated a minimum proficiency equivalent to B2 (i.e., LexTALE score above 60). For the graded condition,[11] the final sample included 74 English L1 monolinguals and 80 German L2 English learners. L2 participants demonstrated a minimum proficiency equivalent to B2 (i.e., LexTALE score above 60; section 2.2.3). Additional participants (16 L1 and 13 L2 participants) were tested but excluded for scoring below 80% correct in optimally true/false *all* trials (pre-registered criterion).

Participants reported no cognitive impairments or learning disabilities and confirmed monolingual upbringing—the latter since our study focuses on L2 development rather than bilingualism and early bilinguals might have implicature derivation advantages over L2 learners (Siegal et al., 2007; Antoniou & Katsos, 2017).

Materials

Twelve scenarios were created, each featuring a set of nine kitchen items, such as apples or mugs (there were always exactly nine apples/mugs/etc.).[12] Each scenario appeared once per sentence type. Throughout the experiment, *some* consistently denoted four out of nine items.

Procedure

Due to Covid-19 restrictions, the experiment was conducted online on *Gorilla* (www.gorilla.sc; Anwyl-Irvine et al., 2020). It took 5–10 minutes.

After giving consent, and before beginning the experiment, participants completed the LexTALE's English version which measures proficiency in a valid and fast way (Lemhöfer & Broersma, 2012; Poort & Rodd, 2019). Participants falling below 60% (B2) could not continue.

In the experiment, participants were introduced to a context where they watch a cooking show on TV together with a foreign friend. The chef moves kitchen items from one place (e.g., shelf) to another (e.g., table). The friend, whose L1 is not English, comments on the chef's every move but occasionally makes language mistakes. During testing, participants saw cartoons of kitchen scenes featuring

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

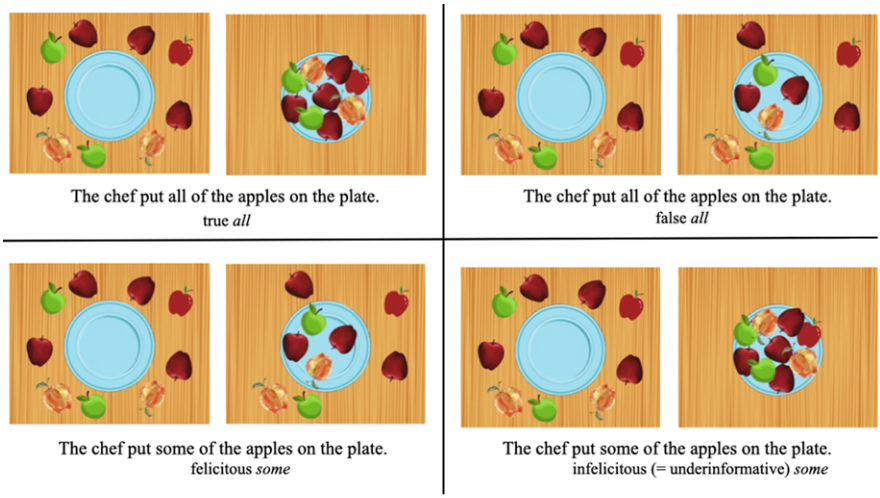

Figure 1. Example of the four sentence types: optimally true *all* (top left), optimally false *all* (top right), felicitous *some* (bottom left), infelicitous (= under-informative) *some* (bottom right). Participants must judge the extent to which the sentences match the picture in each case. Note that responses for the first three types are expected to be highly consistent for proficient speakers (i.e., they will respectively show high levels of: agreeing, disagreeing, and agreeing). For under-informative *some*, high levels of agreement indicate logical responding, whereas low levels of agreement indicate sensitivity to under-informativeness (i.e., that the use of *some* here is under-informative).

target items such as apples or mugs. Each cartoon comprised two pictures (Figure 1). The first picture, always on the left, showed all items in one place (e.g., all apples on the table; all mugs on the shelf). The second picture, always on the right, showed *some*, *none*, or *all* of the initial items in a different place (after the chef had moved them; Figure 1). In each trial, the cartoon was accompanied by a written sentence representing the friend's comment, like *The chef put some/none/all of the apples on the plate*. The sentence and cartoon pictures appeared on the screen simultaneously. Participants had to evaluate whether their friend described the scene correctly using a binary or five-point Likert scale displayed below the sentence. There were 48 trials per participant (twelve scenarios × four sentence types).

To mimic real-life communication situations, participants were asked to react quickly and intuitively. Similarly to Mazzaggio et al. (2021), we introduced 7000 ms time limits per trial. We extended their 3000 ms limit because we provided written and visual input, instead of only aural input. A countdown appeared for the last 5000 ms. The screen proceeded automatically after 7000 ms, regardless of whether a response was provided.

Data analysis plan

Given non-normal data distribution, we conducted non-parametric tests (Wilcoxon signed-rank and Wilcoxon rank-sum tests) to check whether participants from both language groups demonstrate sensitivity to under-informativeness in both tasks. In addition, for both the binary task and the graded task, we evaluate evidence *against*

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

two hypotheses: (i) the hypothesis that L2 users give fewer pragmatic responses to under-informative input and (ii) the hypothesis that L2 users give more pragmatic responses to under-informative input.[13] Note that nonsignificant p-values cannot differentiate between no between-group differences (H0) and no evidence for any conclusion. We therefore used BF as our method of inference which can make this distinction. A BF provides a measure of the extent to which our data support a hypothesized difference between the groups (H1) versus the null that there is no difference between the groups (H0). BF computation therefore requires us to have a model of H1, that is, the distribution of expected differences if H1 is true. Following Dienes (2008, 2014), we compute BFs modeling H1 as a half-normal (thus testing a one-sided prediction, since we are testing a directional hypothesis) with a mean of 0 and the SD set to a rough estimate of the predicted difference if H1 is true. This method requires three numbers: (1) an estimated mean difference in the data; (2) the associated standard error; and (3) an estimate of the predicted mean difference under H1. For (1) and (2), similar to the non-parametric t-test, we use the mean rank difference and associated standard error.[14] For (3), ideally, we would base this on previous data, but there is no study with similar materials using this scale and our pilot did not find a difference we could use, so we lack a relevant value. We used the motivated-maximum approach (cf. Silvey et al. under review and "related room to move" hypothesis in Dienes (2019)). This involves basing a predicted effect size on a maximum and then setting our predicted effect size to twice this value (since our model of H1 is a half-normal with a mean of 0, the maximum is equal to approximate twice the SD). Since we use rank scores, we can compute the logically possible maximum difference between group 1 ($n = n1$) and group 2 ($n = n2$) if no participant in group 2 scored higher than any participant in group 1 (1: n1 ranks are group 1 participants, n1+1:n1+n2 are group 2 participants). This is equal to (n1+n2)/2, that is, in our study: (80+74)/2 = 77 for the graded data and (54+55)/2 = 54.5 for the binary data. We therefore set our rough predicted effect sizes to be equal to half of this value, that is, 38.5 for the graded data and 27.25 for the binary data. It is important to emphasize that while this sets limits on the possible scale of effect, H1 is modeled as a distribution and our use of a half-normal distribution with a mean of 0 corresponds to our expectation that smaller values are more likely.

Although BFs are defined on a continuous scale, we also interpret them according to discrete evidential categories introduced by Jeffreys (1961) and expanded by Dienes (2014) whereby BF>10 indicates strong evidence for H1, BF>3 indicates substantial/moderate evidence for H1, BF<1/3 indicates moderate/substantial evidence for H0, BF<1/10 indicates strong evidence for H1, and otherwise the evidence is ambiguous (i.e., the data are unable to adjudicate between the hypotheses). Since we acknowledge the choice of predicted value to inform H1 is subjective, we also calculated Robustness Regions (RR) for each BF, which show the range of estimates of H1 for which our data would support the same conclusion (i.e., to accept H1/H0 or inconclusive) based on the cutoffs of BF>3 or BF<1/3. This RR is noted as [x1, x2] with x1 being the smallest standard deviation and x2 the largest standard deviation (as recommended by Dienes, 2021). To compute the range, the full range of possible values was tested (i.e., 0 to the maximum of 77 for the graded data and 54.5 for the binary data) in increments of 0.1. Note that larger

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

predicted values bias finding evidence for H0 and smaller values bias finding evidence for H1.

Results

LexTALE

The LexTALE provides a language proficiency measure on a 0% to 100% scale. One L2 participant in the graded task condition scored below 60% (B2) and could not participate. The remaining participants scored: L1 mean = 93.62, SD = 5.71, range 77.5:100; L2 mean = 84.44, SD = 9.81, range 62.5:100. All participants in the binary task condition passed the language test (L1 mean = 94.66, SD = 5.30, range 77.5:100; L2 mean = 86.65, SD = 8.52, range 66.25:100).

Optimally true and optimally false all sentences

Performance on these sentence types was used to filter out low-performing participants. In the graded condition, responses were recoded numerically as follows: optimally true *all*: Disagree = 1, Somewhat Disagree = 2, Neither = 3, Somewhat Agree = 4, Agree = 5; optimally false *all*: Disagree = 5, Somewhat Disagree = 4, Neither = 3, Somewhat Agree = 2, and Agree = 1. In the binary condition, responses were recoded numerically as follows: for optimally true *all*: Disagree = 0, Agree = 1; for optimally false *all*: Disagree = 1; Agree = 0. Participants scoring less than "4" on average on each of the optimally true/false *all* trials in the graded task and participants scoring less than 75% correct answers on optimally true/false *all* trials were removed. In the graded data, 16 L1 and 13 L2 participants were removed. This was unexpectedly high compared with our pilot (conducted in person). On inspection, those participants showed patterns like 100% "disagree" responses to optimally true *all* items, suggesting low levels of attention, a well-known side effect of online research. In the binary data, only one L2 participant was removed. Table 1 shows performance on control items for the remaining sample across both task types.

Main analyses

Analyses were conducted over the under-informative *some* and felicitous *some* data. Since the screen automatically advanced after 7000 ms, 26 responses were missing (less than 1% of the data) for the graded data and three responses (of 2616) for the binary data.

We visualize the data in two ways: first, Figures 2–3 present the proportion of different responses across participants per group for each task type.

On inspection, the data of each task appear to corroborate each other. The response patterns in both tasks look similar for both groups: as expected, for felicitous *some* sentences, most responses are "agree" responses. Critically, however, for the under-informative sentences, most responses in both tasks are "disagree" responses, reflecting (a) a sensitivity to the pragmatic under-informativeness in both groups and (b) an intolerance of such under-informativeness in both groups. Interestingly, the other responses in the graded task are spread, including

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

Table 1. Performance on control items

Task type		Optimally true *all*		Optimally false *all*	
		M	SD	*M*	SD
Quinary	L1 ($n = 74$)	4.97	0.32	4.75	0.73
	L2 ($n = 80$)	4.98	0.18	4.74	0.69
Binary	L1 ($n = 55$)	0.99	0.095	0.97	0.159
	L2 ($n = 54$)	0.99	0.078	0.97	0.173

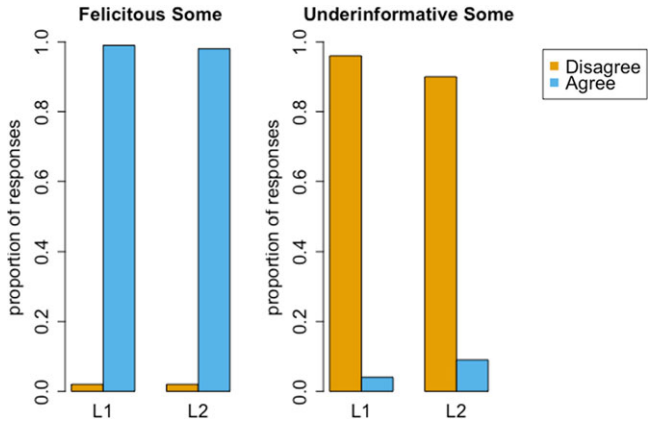

Figure 2. Proportion of different responses across participants for the two types of sentences with *some* (binary tasks).

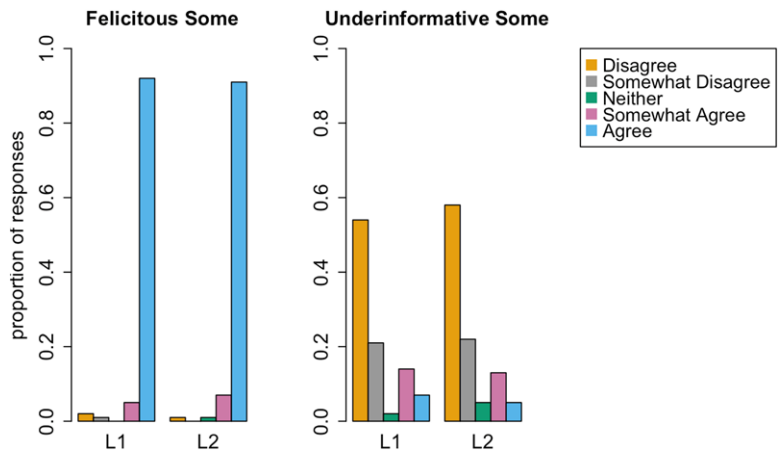

Figure 3. Proportion of different responses across participants for the two types of sentences with *some* (quinary tasks).

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

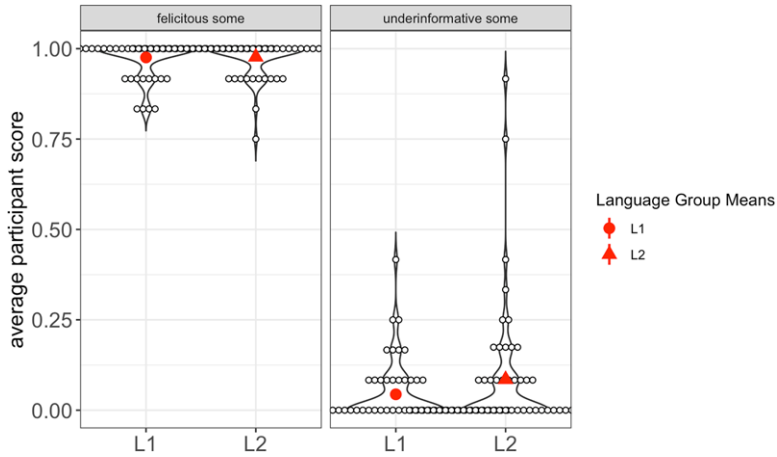

Figure 4. Violin plots showing distribution of participant average scores for each language group on the two types of *some* sentences (binary tasks). Means are shown in red. High scores indicate high levels of agreement that the presented sentences match the pictures.

intermediary responses, suggesting some tolerance of pragmatic violations in both language groups.

Second, we recoded responses numerically and computed participant averages (Figures 4–5). In both tasks, both groups demonstrate similar response patterns. Regarding under-informative items, a large proportion of participants in both groups always answered "disagree" (equaling an average of 1 or 0), with other participants having average scores which indicate that they gave more mixed responses, though leaning toward the scale's "disagree" end (no participant in either group always "agreed" which would be an average of 5 or 1). In fact, the binary data suggest that only two L2 participants show an overall tendency to tolerate pragmatic violations (i.e., average score $> .5$). However, the graded data indicate that without the pressure of having to make a dichotomous choice, a number of participants in both language groups tend to tolerate (i.e., average score > 3) the pragmatic violations.

The visualizations demonstrate that both groups show strong tendencies to disagree with under-informative *some* sentences in contexts where *all* would be more appropriate, regardless of the task type. Nonetheless, the more nuanced graded data indicate that if given the option to participants in both language groups tend to be willing to show some tolerance of pragmatic violations.

Turning to our statistical analyses, these were conducted over participant average scores (Figures 4–5). Since the data are not normally distributed, we employed non-parametric tests. Our first hypothesis is both language groups are sensitive to under-informativeness. To test this, we tested with Wilcoxon signed-rank tests whether mean response scores for control items (felicitous *some*) were significantly lower than those for target items (under-informative *some*). Since this logic applies to both task types, the tests were conducted for both task types. Results demonstrate that mean response scores to under-informative *some* were lower than those to felicitous *some* in both language groups, confirming our first hypothesis (binary task: L1:

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

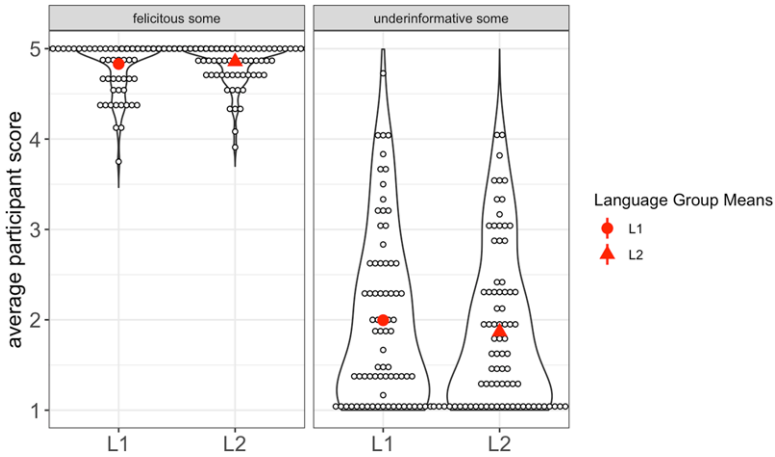

Figure 5. Violin plots showing distribution of participant average scores for each language group on the two types of some sentences (quinary tasks). Means are shown in red. High scores indicate high levels of agreement that the presented sentences match the pictures.

$p<.001$, $r=.89$ // L2: $p<.001$, $r=.89$; graded task: L1: $p<.001$, $r=.87$; L2: $p<.001$, $r=.87$).

Our second and key question concerns the between-group differences. For both task types, we began by computing difference scores between felicitous *some* and under-informative *some* for each participant and conducted Wilcoxon rank-sum tests comparing both groups.[15] This suggested no significant between-group difference ($p > .05$). Given the shortcomings of the p-value statistic (see *Data Analysis Plan*), we calculated BFs.

For the binary data, we used the same difference scores used to compute the rank-sum test and computed BFs to test each of the hypotheses that (1) *English L2 speakers give fewer pragmatic responses to under-informative sentences than English L1 speakers* (which predicts lesser rejection of under-informative *some* and thus smaller difference scores for L2 participants) and (2) *English L2 speakers give more pragmatic responses than English L1 speakers* (which predicts greater rejection of under-informative *some* and thus larger difference scores for L2 participants). Recall that although we do not anticipate group differences, we must test the null hypothesis against the positive hypothesis that such differences exist. Starting with 1, using the method described above, we required a mean difference (in rank scores) between the groups (mean = 6.99) and standard error (SE = 5.56). Computing and comparing this data summary's probability under H0 versus under H1 where H1 is modeled as a half-normal with a mean of 0 and a SD set to our rough estimate of predicted difference (i.e., 27.25) resulted in BF = 0.76. As BF>1/3, we conclude that our evidence is ambiguous; hence, we cannot determine from this data whether H0 or H1 is more likely. For 2, the hypothesis in the opposite direction, our data model (mean = −6.99; SE = 5.56; predicted difference = 27.25) resulted in BF = 0.095. We take this as evidence for H0. RRs were RR[7.0:∞] indicating that we could have used a four times smaller estimate of the predicted effect and still found evidence for H0 (BF<1/3).[16]

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

For the graded data, we again used the same difference scores used to compute the rank-sum test and computed BFs to test each of the hypotheses that (1) *English L2 speakers give fewer pragmatic responses to under-informative sentences than English L1 speakers* and (2) *English L2 speakers give more pragmatic responses than English L1 speakers*. For (1), our data model (mean = −4.69; SE = 7.19; predicted difference = 38.5) resulted in BF = 0.12. We take this as evidence for H0. RRs were RR[12.5:∞] indicating that we could have used a three times smaller estimate of the predicted effect and still found evidence for H0. For (2), our data model (mean = 4.69; SE = 7.19; predicted difference = 38.5) resulted in BF = 0.33. We take this as evidence for H0. RRs were RR[38.32:∞].

Discussion

This study investigates L2 and L1 users' response behavior toward under-informative input using binary and graded TVJ tasks. In the binary task, L1 and L2 participants overwhelmingly showed pragmatic responses when "forced" to make dichotomous judgments. In Katsos and Bishop's (2011) terms, both groups showed intolerance of pragmatic violations. In the graded task, the same tendency emerged; however, the spread of responses revealed some tolerance of the pragmatic violation in both groups. (Or alternatively, some recognition that there was an alternative logical interpretation, and sensitivity to that). We did not find evidence of between-group differences in either experiment. Additional Bayesian analyses with the binary data provided evidence against the hypothesis that L2 users provide more pragmatic responses to under-informative sentences than L1 users, as reported by Slabakova (2010). Evidence against the hypothesis that L2 users provide fewer pragmatic responses than L1 users (Mazzaggio et al., 2021) was ambiguous. Bayesian analyses with the graded data provided evidence against both the hypotheses that L2 users provide more pragmatic responses to under-informative sentences than L1 users and that L2 users provide fewer pragmatic responses than L1 users. These findings contradict claims about between-group differences in previous L2 work and help reshape our perception of L2 learners' pragmatic abilities compared to L1 users.

Sensitivity to under-informativeness

Regarding fundamental pragmatic abilities, our data show that both language groups are sensitive to under-informativeness, and that this sensitivity did not differ between groups. This is unsurprising. As outlined above, we would not expect developmental differences between adult L1 and L2 users that might impact such fundamental pragmatic awareness, nor would we anticipate the "cognitive burden" of considering the potential for a more informative statement to exceed the cognitive capacity of L2 users. Regarding implicature derivation, it is highly likely that participants were able to derive the implicature, thereby accessing the pragmatic interpretation of some (*some but not all*), as indicated by ceiling performances in the felicitous *some* trials. Here, agreement only makes sense when one has access to the *some but not all* interpretation. However, for the reasons outlined above, our TVJ data, specifically data from under-informative trials, cannot fully speak to the issue of implicature derivation. Even if we assumed based on the

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

ceiling responses in felicitous trials that participants had access to the pragmatic interpretation, data from TVJ tasks do not reveal processing order or time course although this information is crucial to adjudicate between different theoretical accounts of implicature derivation. Such information would require other experimental methods like eye-tracking or ERP techniques that provide insights into different implicature derivation stages (Huang & Snedeker, 2009; Yoon et al., 2015). For example, Noveck and Posada's (2003) ERP data indicated no immediate cognitive reaction to under-informative sentences, which allowed them to make assumptions about time courses and triggers of implicature processing. Similar L2 research might reveal similar processing details that go beyond sensitivity to under-informativeness, allowing nuanced theoretical conclusions.

L1–L2 differences

As discussed in 1.3, only Dupuy et al. (2019) reported null results, although their statistics were inappropriate for evaluating evidence for H0. In contrast, Slabakova (2010), Lin (2016), and Snape and Hosoi (2018) found that L2 participants were more likely to reject under-informative input than L1 participants (though results were only significant in the first two studies), whereas Mazzaggio et al. (2021) report the opposite. Why do we not see the same patterns reported in previous studies?

The experimental pragmatics literature offers some explanations as to which factors might "push" participants to accept or reject pragmatic violations. For example, Feeney and Bonnefon (2012) found effects of politeness contexts and individuals' honesty traits on the interpretation of scalar expressions (see *Previous L2 studies*). Assumptions of honesty were influential in a study by Sikos et al.'s (2019) as well. They manipulated speakers' social attributes (i.e., likeability and L2 status) and found that those manipulations influenced acceptability ratings of under-informative items in binary tasks. If speakers were perceived as likeable, participants were more likely to accept under-informative utterances, and, vice versa, if speakers were perceived less likeable, participants were less likely to accept under-informative items. Interestingly, made-up "non-native speakers" were rated lowest by the participants (who were native speakers) in terms of likeability, and their under-informative utterances were most likely to be penalized. Importantly, replicating the same experiment using a graded task, Sikos et al. (2019: 9-10) find that the "correlation between speaker likeability and acceptance of critical items is eliminated" and conclude that "when forced to select between two inapt options in a binary choice task, social factors can tip the balance so that participants choose to reject [under-informative] statements more often for certain speakers." This demonstrates how social factors independent of linguistic influences might impact decision-making processes in binary tasks. Turning to L2 studies, social influences like those influential in Feeney and Bonnefon's (2012) and Sikos et al.'s (2019) studies might have been at play in Slabakova's (2010) second experiment where pragmatic responses in both groups increased compared to the first experiment. Unlike the first experiment where participants judged non-contextual sentences drawing on encyclopedic knowledge ("Some elephants have trunks"), the second experiment introduced picture-stories featuring real humans where the context implied that the speaker attempted to deceive the hearer through

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

their under-informative use of *some*. As noted by Dupuy et al. (2019), the pragmatic interpretation becomes more relevant in such scenarios, because that is what the speaker is intending to convey.

While factors other than cognitive factors may impact decision-making in the context of judging pragmatic violations, the question remains why Slabakova (2010) found a pragmatic "advantage" among L2 users compared to L1 users. Mazzaggio et al. (2021) suggest that the immersion students recruited for Slabakova's study might have had a cognitive advantage over the L1 users, resulting in higher implicature derivation rates. However, this argument does not apply when we assume that the study's binary tasks might not have captured implicature derivation but rather display the trade-off between two competing interpretations of *some*. In addition, ceiling performances in felicitous control tasks across both language groups in Slabakova's (2010) study indicate that both groups could access the pragmatic interpretation. Another possibility might be that the L2 pragmatic bias in L2 users was the result of their L1 background. In contrast to our own and other L2 studies (Dupuy et al., 2019; Mazzaggio et al., 2021) featuring linguistically and culturally relatively similar languages (German, Italian, Spanish, French, and English), Slabakova's (2010) study featured English and Korean, two relatively "distant" linguistic and cultural backgrounds. Perhaps, there are linguistically and culturally engrained tendencies toward accepting or rejecting utterances perceived as "wrong" or "untruthful." In addition, cross-linguistic differences like the perceptions of what fraction of a quantity felicitously constitutes "some" (Stateva et al., 2019) could influence decision-making at the intercept of logical meaning, pragmatic meaning, and social context. This interplay of quantifier meaning and social context is demonstrated by Zhang and Wu (2020) who report that Chinese speakers employ different interpretations of *some* depending on whether the quantifier is used as informative (pragmatic interpretation) or polite item (logical interpretation). Thus, it seems plausible that *pragmalinguistic failure* which "occurs when the pragmatic force mapped by [L2 users] onto a given utterance is systematically different from the force most frequently assigned to it by native speakers of the target language" (Thomas, 1983: 99) might cause different response patterns across L1 and L2 groups. Yet, the available evidence seems to speak against this view of fundamental group differences caused by cultural and linguistic backgrounds: when tested in their L1, the Korean speakers' response behavior did not differ to English L1 users in Slabakova's (2010) study. The same appears to be true for Snape and Hosoi's (2018) Japanese and English speakers, as indicated by their descriptive data.

While the null results of our binary and graded tasks correspond to our expectations (based on using two relatively similar language groups and the absence of manipulations of social context), it would nevertheless be interesting to see whether recruiting culturally and linguistically vastly different samples and manipulating social contexts would impact participants' response patterns in both binary and graded tasks and potentially cause group differences. Note, such investigations might no longer be a question of whether participants are completing the experiment in their L2 but instead depend on general L1-specific factors. For example, future studies could manipulate speakers' honesty traits which might provoke differing pragmatic tolerance across groups with vastly different L1

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

backgrounds. This could be achieved by creating a scenario where a hearer is cooking and cannot see the TV in the living room, yet s/he requires information about the TV chef's actions and relies on the speaker's report. Prior to testing, the speaker's honesty traits could be manipulated ("S/he misguides people for a laugh"; "S/he is very honest") and participants are assigned to different "honesty" and binary/graded conditions.

We now turn to Mazzaggio et al. (2021) who report results in the opposite direction. Their participants gave *fewer* pragmatic responses in L2 than L1. We did *not* find evidence for this in either experiment. When we use the same binary scale as Mazzaggio et al., evidence for H0 is inconclusive. It trended in the direction of more evidence for H0 but did not cross our pre-registered boundary of $BF < 1/3$ for substantial evidence. In the graded task, we found evidence for H0 that L2 users did not provide fewer pragmatic responses than L1 users. Taking both experiments' results together, while we can conclude that L2 participants are *not* less sensitive to pragmatic violations than L1 speakers, we cannot rule out that they are nevertheless more willing to be tolerant of such violations when given a binary scale. On inspection of the data, this ambiguity seems driven by the fact that two L2 participants give more "agree" than "disagree" responses. To resolve the group differences question, it would be necessary to recruit further participants. Here, we stopped at our pre-registered maximum (which was set based on available resources).

Given the inconclusive BF, we cannot say that our data are incompatible with Mazzaggio et al.'s (2021). Still, we tentatively consider why we don't find evidence for the positive effect in that paper: in addition to the potentially relevant contextual factors discussed above, the literature offers several other relevant factors which influence the decision-making in the context of pragmatic violations, including the effects of the exact type of wording used to elicit responses (*Agree/Disagree; Yes/No; That's (not) right*; Mazzaggio et al., 2021; Kissine & de Brabanter, 2023), the partitive quantifier formulations (*some X* vs. *some of the X*; Degen & Tanenhaus, 2013), or the discourse context (*question under discussion* (QUD), Yang et al., 2018; Dupuy et al., 2016). On the one hand, those factors might have mediated the impact of language background in previous L2 studies, which could explain the inconsistent results. On the other hand, those factors have varied in an unsystematic manner across all the previous L2 studies and are not specific to Mazzaggio et al. (2021). What is exclusive to Mazzaggio et al.'s (2021) study is the use of aural input. While they introduced aural input to increase the L2 users' cognitive burden with the aim of adjudicating between different theoretical accounts of implicature derivation, the aural nature of the input might have had a different effect, caused by prosody. The experimental pragmatics literature suggests that prosody can impact the rate of pragmatic inferences (cf. Reboul & Stateva, 2019). For example, Chen et al. (2018) demonstrated with Chinese L1 users that prosodic cues (i.e., stress on individual words) can help listeners arrive at the inference intended by the speaker (cf. Bill et al., 2018, experiment 2). It makes sense that intonation impacts comprehension, and it might potentially be the case that Mazzaggio et al.'s (2021) recording (unintentionally) featured relevant prosodic cues. Yet, it remains unclear why such cues would impact one but not the other language group. Regarding the differences between Mazzaggio et al.'s (2021) and our current work, what seems more plausible

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

is that their tight time limit (3000 ms) in combination with the aural nature of their input might have impacted response patterns. Perhaps their L2 users did in fact require more time to process the sentences. For example, it might have taken them longer to decode the phonological input. Even just a few additional milliseconds of processing take time away from the taxing decision-making process over which interpretation to give preference to. While this view does not explain the direction of the L1–L2 differences reported by Mazzaggio et al. (2021), it might contribute to an account of why there might have been differences in the first place. And although our study featured a time limit as well, it was considerably less pressing (7000 ms), potentially allowing both language groups enough time to arrive at their decision without the need to rush it.

Comparison of binary and quinary data

Our results contrast Katsos and Bishop's (2011) in that they do not suggest large differences in participants' pragmatic responding when given a binary versus graded scale. Most likely, children's tolerance of pragmatic violations when given binary choices in previous studies is an age-based developmental factor which does not apply in our study. Nonetheless, we see qualitative differences in our two experiments' response patterns. In the graded task, while many individual participants give 100% "disagree" responses, many others provide intermediate responses including some participants whose responses suggest that on average they are using the "agree" end of the scale (i.e., Figure 5 includes participants who are above 3, the scale's halfway point). Similar to Sikos et al.'s (2019) study, our data suggest that when given a non-dichotomous choice participants demonstrate some tolerance of pragmatic violations. In contrast in the binary task, a larger proportion of participants give 100% "disagree" responses and, while there is still some evidence of intermediary responding across trials (i.e., participants who do not give 100% "disagree" responses), there are only two (L2) participants whose responses indicate they mostly "agreed" (i.e., above 0.5 in Figure 4).

Given this pattern, we conducted some additional (non pre-registered) analyses to explore statistically whether these qualitative differences hold if the quinary responses' spread—for comparability—is translated into the assumptions of "traditional" binary coding. We use a coding inspired by Jasbi et al. (2019). In their study, they suggest two linking hypotheses regarding assumptions about how responses on graded scales link to binary scales. (They assume that with binary data, "agree"/"disagree" responses indicate not-deriving/deriving implicatures, respectively). They differentiate a "strong" linking hypothesis, which assumes that only full "disagree" responses indicate implicature derivation (while any other response indicates that it was not) and a "weak" linking hypothesis which assumes that implicatures are derived when any of "disagree," "somewhat disagree," "neither," or "somewhat agree" are chosen. We explored the implications of applying these two types of coding in our own data, re-coding responses using the "weak" and "strong" criteria (Figures 6–7). For the "strong" and "weak" re-codings separately, we fitted mixed logit models with a random intercept by participant. The models revealed a significant effect of task type on response behavior only under the strong interpretation (weak link: $\beta = 0.10$, $SE = 0.26$, $z = 0.39$, $p > .05$; strong-link: $\beta = 4.71$,

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

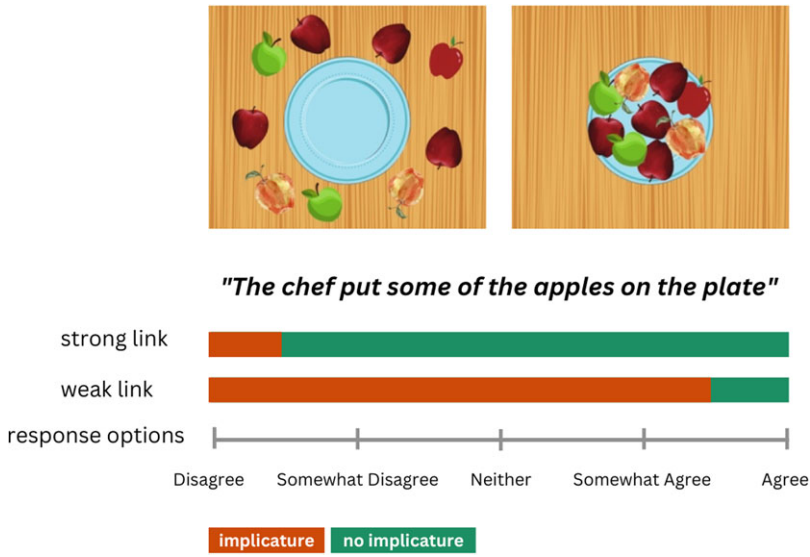

Figure 6. Strong and weak link from response options to researcher inference about scalar implicature rate, exemplified for under-informative *some* when the quantifier *all* would be more informative (adapted from Jasbi et al., 2019).

Figure 7. Inferred implicature rates on under-informative *some* trials (see Figure 6) as obtained with the binary and quinary judgment task. Y-axis represents the proportion of total responses which would be considered as "implicature derived" or "implicature not derived" depending on the applied linking hypothesis (strong/weak) (cf. Jasbi et al., 2019).

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

SE = 0.52, z = 8.99, p < .05). In terms of Jasbi et al.'s criteria, this is in the direction of suggesting more implicature derivation in the binary task. This interpretation is unintuitive: it seems unlikely that the less nuanced scale would be more rather than less revealing as to pragmatic competence. Instead, we suggest the difference seen in the strong coding underlines the point we made above (cf. Sikos et al., 2019)—that there is a larger proportion of full "disagree" responses in the binary data, which does not reflect differences in pragmatic "competence" (i.e., whether participants derive implicatures) but rather in participants' willingness to accept violations. Nevertheless, this underlines the importance of being clear in our assumptions when interpreting results of pragmatic studies.

The interplay of "felicitous some" and "infelicitous some" trials

Although not part of our main inquiry, we now discuss the often-overlooked interplay between felicitous and infelicitous *some* trials. As noted earlier, only the *some but not all* interpretation licenses agreement in felicitous trials. The fact that participants consistently perform at ceiling in the felicitous trials in our study and elsewhere (Dupuy et al., 2019; Slabakova, 2010; Mazzaggio et al., 2021) seems to suggest that they can generally access the pragmatic interpretation of *some*, irrespective of processing nuances within and between groups. The ceiling performances also seem to suggest that the impact of the logical reading (*some and possibly all*) is minimal in felicitous trials. This makes sense: based on their linguistic experience (Kissine & de Brabanter, 2023) and an expectation of cooperation (Grice, 1975), participants are likely to assume that if speakers use the term "some" they aim to convey the pragmatic meaning. Reversely, regarding infelicitous trials, assuming participants can access the pragmatic interpretation of *some*, we would expect rejections. Yet, across studies and task types, participants do not consistently reject infelicitous input. So far, we argued that agreement in binary tasks or, similarly, the response spread in graded tasks result from the balancing act between sensitivity to violations and the tolerance thereof. However, we suggest that perhaps any response other than extreme "reject" responses indicate whether participants could access the *logical* interpretation *some and possibly all*. Pilot data of the current study yielded anecdotal participant feedback suggesting that they took multiple infelicitous trials to realize that there is a logical "reading" of *some* available, and some participants failed entirely to entertain the logical reading. Both behaviors are understandable because participants' linguistic experience and their expectation of cooperativity strongly speak against expecting *some* being used in a logical manner. Thus, perhaps it is not the pragmatic but rather the logical interpretation which is not always readily available to participants in TVJ tasks. In fact, if or once available, it might even be considered odd as it is extremely unexpected. Therefore, we wonder whether among all the inquiries into participants' access to the *pragmatic* interpretation (and its derivation), we overlooked that what might rather be under scrutiny in infelicitous trials is participants' access to the *logical* interpretation.

Since this notion is speculative and our data cannot speak to it, we propose future research could investigate the following aspects: first, *if* and *when* (i.e., number of infelicitous trials) participants entertain the logical interpretation throughout an experiment like ours. Second, considering that participants overwhelmingly seem to

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

demonstrate access to the pragmatic interpretation through their agreement in felicitous trials (although we cannot be certain), we wonder if response behaviors gradually change across infelicitous trials, specifically once participants start considering the logical interpretation. From this moment, their decision-making might not only depend on their awareness and tolerance of pragmatic violations anymore but might additionally be mediated by the availability of the logical interpretation. (Access to the logical interpretation might even impact behavior in felicitous trials because it presents a viable interpretation option). This also raises the question whether explicitly alerting participants to the logical interpretation before the experiment changes their response behaviors across trial types.

The inquiries regarding the availability of the logical interpretation might further complicate matters regarding TVJ tasks. The general question remains, what do they measure *exactly*? Our data only show that L1 and L2 users consistently notice that some sentences are under-informative, and that the two groups do not differ in their binary and graded response behaviors. We discussed potential reasons why our data and the data of previous L2 studies might differ from each other. Yet, TVJ tasks reveal little about underlying factors determining participants' decision-making process, and the TVJ data yielded by our study and others may be insufficiently informative for exploring them.

Limitations and future directions

A potential limitation is that we worked with a particular L2 group—German speakers—without testing them in their L1. Such controls are reported elsewhere (Slabakova, 2010) and are important in ruling out that between-group differences are not due to cross-linguistic differences, like in quantifier interpretation (Stateva et al., 2019). However, as we found no evidence of group differences, this seems less critical here.

Although we consider using BFs as an advantage of this study, we acknowledge that there is subjectivity in priors which we use as our H1 model. Dienes 2008; 2019 emphasizes the importance of using priors with parameters in the right ballpark for the theory in hand. Ideally, we would have used values from previous data, but unfortunately for our graded data analyses, we lacked appropriate data on the same scale; thus, we used a scale factor for the prior working backward from a maximum based on the scale ("motivated-maximum approach"; Silvey et al. 2021). While this is better than using uninformed defaults, we acknowledge that it is potentially biased in the direction of overestimating the potential effect size (thus biasing H0). To mitigate against this, we provided sensitivity analyses showing the extent to which our findings would hold given different priors. Moreover, with the binary data we had an available value from previous work to inform the H1 prior, and we compared analyses using this prior (with estimates based on logistic regression) to analyses with a prior based on the motivated-maximum approach (same approach as graded data). Reassuringly, we found qualitatively the same results when testing each of the directional hypotheses (i.e., evidence in favor of H0 for the hypothesis that L2 participants would give more pragmatic responses, ambiguous evidence for the hypothesis that they would give less pragmatic responses). This gives us some confidence in the motivated-maximum approach. In allowing us to quantify

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

evidence for H0, BFs allow us to draw stronger conclusions about the "lack of" group differences compared with previous studies with null results in this respect. We acknowledge that with the binary data, BF analyses revealed that we *did not* have a sufficient sample to rule out one of our hypotheses. We considered this in the Discussion.

Conclusion

This study demonstrates that both proficient L2 English learners and English monolinguals are sensitive to under-informativeness in contexts with under-informative *some*. In contrast to previous research, we found no evidence of between-group differences in the extent of pragmatic responses when using either binary or quinary tasks. Moreover, for the quinary task, we found evidence that there was *no* between-group difference either in the direction of more pragmatic responding in L2 or less pragmatic responding in L2. Regarding individual participants' response patterns, in both groups we found some differences with the binary and quinary scale. This was in the direction of apparently more pragmatic (less logical) responding in the binary task, which we interpret in terms of metalinguistic attitudes pushing participants to reject pragmatic violations in this particular task and context.

Replication Package. The pre-registration as well as study materials, analysis scripts, and anonymized data are available on https://osf.io/6bt53/.

Acknowledgements. We would like to thank the three anonymous reviewers for their detailed and helpful feedback on earlier versions of this manuscript. We also thank the participants in our pilot and main study.

Competing interests. The authors have no conflicts of interest to declare.

Notes

1 They also tested the *Semantic Error Hypothesis* which falls outside the scope of the current study.

2 Kissine and de Brabanter (2023) further suggest that rejections of under-informative sentences might not even require participants to explicitly activate the stronger alternative (assumed by Katsos & Bishop, 2011). Instead, it might be sufficient to realize that the sentence is infelicitous based on the experience that cooperative speakers would never use *some X are Y* in scenarios where obviously *all X are Y*. However, their experimental data cannot speak to this issue.

3 Since standard frequentist statistics were used, the study could not confirm the absence of group differences in the binary data.

4 We thank an anonymous reviewer for making this point.

5 Common European Framework of Reference for language levels.

6 The pre-registration as well as study materials, analysis scripts, and anonymized data are available on https://osf.io/6bt53/.

7 If one were oblivious to the pragmatic violation, a rejection would not make sense.

8 Note the graded task is important for another reason. L2 users' responses in binary tasks in previous L2 research provided evidence in "both directions" (i.e., tendency towards rejection and tendency towards acceptance). If participants overwhelmingly accept pragmatic violations in the binary task, their sensitivity to under-informativeness would be masked. In this case, the more nuanced graded task might help distinguish (a) sensitivity to under-informativeness simply overridden by pragmatic tolerance from (b) potential obliviousness of pragmatic violations. The former would be indicated by any response but extreme "Agree" and the latter by extreme "Agree" responses only (cf. Veenstra et al., 2018, Table 1).

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

9 Being a result of revisions, this slightly deviates from our pre-registration where we suggested one-tailed testing in one direction only.
10 Identical patterns of results are found when conducting equivalent analyses directly on scores from under-informative items. We report results with difference score analyses here (since this was pre-registered). Additional analyses are available in the analysis script.
11 Power simulations determined sample sizes for both experiments. Following a setup error, we recruited more participants for the graded condition than required.
12 Although other studies used smaller sets (Dupuy et al., 2019), Snape and Hosoi (2018) noted that small sets make scalar scales inconclusive and impact interpretation.
13 For the graded task, we thank a reviewer for their suggestion to conduct additional CLMM analyses more appropriate for ordinal data (cf. Christensen, 2023). The results did not differ qualitatively from the results reported in the main text. The additional analyses are available in the Appendix as part of the analysis script.
14 We obtained this by running a t-test over the ranks and obtaining the standard error by dividing the mean difference by the t. (Note the non-parametric t-test is essentially a t-test on the ranks; Zimmerman and Zumbo, 1993).
15 We ran two-sided and one-sided tests.
16 Additional analyses on the binary data using logistic mixed-effect modelling (like Mazzaggio et al., 2021) yielding the same qualitative outcome are provided in the script.

R Packages

lme4

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. doi: 10.18637/jss.v067.i01

lmerTest

Kuznetsova, A., Brockhoff, P.B. & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed EffectsModels. *Journal of Statistical Software*, 82(13), 1-26. doi: 10.18637/jss.v082.i13

ordinal

Christensen, R. H. B. (2022). ordinal - Regression Models for Ordinal Data. R package version 2022. 11-16. https://CRAN.R-project.org/package = ordinal.

RColorBrewer

Neuwirth, E. (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package = RColorBrewer

rstatix

Kassambara, A. (2021). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.0. https://CRAN.R-project.org/package = rstatix

tidyverse

Wickham et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. doi: 10.21105/joss.016.

References

Antoniou, K., & Katsos, N. (2017). The effect of childhood multilingualism and bilectalism on implicature understanding. *Applied Psycholinguistics*, **38** (4), 787–833.

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. (2020). Gorilla in our midst: an online behavioral experiment builder. *Behaviour Research Methods*, **52**, 388–407.

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: the role of scalar alternatives in children's pragmatic inference. *Cognition*, **118**, 87–96.

Bill, C., Romoli, J. & Schwarz, F. (2018). Processing presuppositions and implicatures: similarities and differences. *Frontiers in Communication*, **3**, Article 44.

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.

Bouton, L. F. (1992). The interpretation of implicature in English by NNS: does it come automatically – without being explicitly taught? *Pragmatics and Language Learning*, **3**, 53–65.

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

Carston, R. (1998). Informativeness, relevance and scalar implicature. In R. Carston & S. Uchida (Eds.), *Relevance theory: applications and implications* (pp. 179–236). John Benjamins.

Chen, I., Huang, C. & Politzer-Ahles, S. (2018). Determining the types of contrasts: the influences of prosody on pragmatic inferences. *Frontiers in Psychology*, **9**, Article 2110.

Christensen, R. H. B. (2023). A Tutorial on fitting Cumulative Link Mixed Models with clmm2 from the ordinal Package. https://cran.r-project.org/web/packages/ordinal/vignettes/clmm2_tutorial.pdf.

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

Davies, C. & Katsos, N. (2010). Over-informative children: production/comprehension asymmetry or tolerance to pragmatic violations? *Lingua* **120**(8), 1956–1972.

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. (2013). Making inferences: the case of scalar implicature processingtle. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual meeting of the cognitive science society* (pp. 3299–3304). Cognitive Science Society.

Dienes, Z. (2008) *Understanding psychology as a science: an introduction to scientific and statistical inference*. Palgrave Macmillan.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, **5**, 781.

Dienes, Z. (2019). How do I know what my theory predicts? *Advances in Methods and Practices in Psychological Science*, **2**(4), 364–377.

Dienes, Z. (2021). How to use and report Bayesian hypothesis tests. *Psychology of Consciousness: Theory, Research, and Practice*, **8**(1), 9–26.

Dupuy, L., Stateva, P., Andreetta, S., Cheylus, A., Déprez, V., Henst, J. B. V. D., Jayez J., Stepanov A., & Reboul, A. (2019). Pragmatic abilities in bilinguals: the case of scalar implicatures. *Linguistic Approaches to Bilingualism*, **9**(2), 314–340.

Dupuy, L., Van der Henst, J., Cheylus, A. & Reboul, A. C. (2016). Context in Generalized Conversational Implicatures: the Case of Some. *Frontiers in Psychology*, **7**, Article 381.

Feeney, A., Scrafton, S., Duckworth, A. & Handley, S. J. (2004). The story of some: everyday pragmatic inference by children and adults. *Can. J. Exp. Psychol.*, **58**(2), 121–132.

Feeney, A. & Bonnefon, J.-F. (2012). Politeness and honesty contribute additively to the interpretation of scalar expressions. *Journal of Language and Social Psychology*, **20**(10), 1–10.

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

Grice, H. P. (1975) Logic and conversation. In P. Cole & H. Morgan (Eds.), *Syntax and semantics. Vol. 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**(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.

Horn, L. R. (1972). On the Semantic Properties of Logical Operators in English (Doctoral dissertation). Retrieved from https://linguistics.ucla.edu/images/stories/Horn.1972.pdf.

Horn, L. R. (2004). Implicature. In L. R. Horn & G. Ward (Eds.), *The Handbook of Pragmatics* (pp. 3–28). Blackwell Publishing.

Horn, L. R., Allan, K., & Jaszczolt, K. M. (2012). Implying and inferring. In K. Allan & K. M. Jaszczolt (Eds.), *The Cambridge Handbook of Pragmatics* (pp. 69–86). CUP.

Huang, Y. T. & Snedeker, J. (2009). Semantic meaning and pragmatic interpretation in 5-year-olds: evidence from real-time spoken language comprehension. *Developmental Psychology*, **45**(6), 1723–1739.

Jasbi, M., Waldon, B., & Degen, J. (2019). Linking hypothesis and number of response options modulate inferred scalar implicature rate. *Frontiers in Psychology*, **10**, 189.

Jeffreys, H. (1961). *Theory of probability* (3 ed.). Oxford: Oxford University Press.

Katsos, N. & Bishop, D. (2008). Pragmatic Tolerance. Paper presented at the *XI International Congress for the Study of Child Language (IASCL)*, Edinburgh, UK, 28 July–1 August.

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

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

Katsos, N. & Smith, N. (2010). Pragmatic tolerance and speaker-comprehender asymmetries. In K. Franich, K. M. Iserman & L. L. Keil (Eds.), *The 34th Boston university conference in language sevelopment – proceedings* (pp. 221–232). Cascadilla Press.

Katsos, N. (2009). Evaluating under-informative utterances with context-dependent and context-independent scales: experimental and theoretical implications. In U. Sauerland & K. Yatsushiro (Eds.), *Semantics and pragmatics: from experiment to theory* (pp. 51–73). Palgrave Macmillan.

Thomas, J. (1983). Cross-cultural pragmatic failure. *Applied Linguistics*, **4**(2), 91–112.

Khorsheed, A., Rashid, S., Nimehchisalem, V., Geok Imm, L., Price, J. & Ronderos, C. R. (2022). What second-language speakers can tell us about pragmatic processing. *PLoS ONE*, **17**(2), e0263724.

Kissine, M. & De Brabanter, P. (2023). Pragmatic responses to under-informative some-statements are not scalar implicatures. *Cognition*, **237**, 105463.

Lemhöfer K. & Broersma, M. (2012). Introducing LexTALE: a quick and valid lexical test for advanced learners of English. *Behaviour Research Methods*, **44**, 325–343.

Levinson, S. C. (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), e0145494.

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

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

Noveck, I. & Sperber, D. (2007). The why and how of experimental pragmatics: the case of 'scalar inferences.' In N. Burton-Roberts (ED.), *Pragmatics* (pp. 307–330). Palgrave.

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

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

Papafragou, A. & Tantalou, N. (2004). Children's computation of implicatures. *Language Acquisition*, **12**(1), 71–82.

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

Poort, E. D. & Rodd, J. M. (2019). Towards a distributed connectionist account of cognates and interlingual homographs: evidence from semantic relatedness tasks. *PeerJournal*, **7**, e6725.

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.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reboul, A. C. & Stateva, P. (2019). Editorial: scalar implicatures. *Frontiers in Psychology*, **10**, Article 1767.

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

Sauerland, U. (2012). The computation of scalar implicatures: pragmatic, lexical or grammatical? *Language and Linguistics Compass*, **6**(1), 36–49.

Schmitt, C., & Miller, K. (2010). Using comprehension methods in language acquisition research. In E. Blom & S. Unsworth (Eds.), *Experimental methods in language acquisition research* (pp. 35–56). John Benjamins.

Siegal, M., Matsuo, A. & Pond, C. (2007). Bilingualism and cognitive development: Evidence from scalar implicatures. In Y. Otsu (Ed.), *Proceedings of the Eighth Tokyo Conference on Psycholinguistics* (pp. 265–280). Hituzi Syobo.

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

Silvey, C., Dienes, Z. & Wonnacott, E. (2021). *Bayes factors for mixed-effects models*. PsyArXiv. doi: 10.3389/fpsyg.2019.00510

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**(2), 163–192.

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press

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

Stateva, P., Stepanov, A., Déprez, V., Dupuy, L. E. & Reboul, A. C. (2019). Cross-linguistic variation in the meaning of quantifiers: implications for pragmatic enrichment. *Frontiers in Psychology*, **10**, 957.

Veenstra, A. & Katsos, N. (2018). Assessing the comprehension of pragmatic language: sentence judgment tasks. In A. H. Jucker, K. P. Schneider, & W. Bublitz (Eds.), *Methods in Pragmatics* (pp. 257–279). de Gruyter Mouton.

Veenstra, A., Hollebrandse, B. & Katsos, N. (2018). Why some children accept under-informative utterances. *Pragmatics & Cognition*, **24**(2), 297–314.

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

Yoon, E. J., Wu, Y. C., & Frank, M. C. (2015). Children's online processing of ad-hoc implicatures. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th annual conference of the cognitive science society* (pp. 2757–2762). Cognitive Science Society.

Zhang, J. & Wu, Y. (2020). Only *youxie* think it is a nice thing to say: interpreting scalar items in face-threatening contexts by native Chinese speakers. *Journal of Pragmatics*, **168**, 19–35.

Zimmerman, D. & Zumbo, B. (1993). Significance testing of correlation using scores, ranks, and modified ranks. *Educational and Psychological Measurement*, **53**, 897–904.

Ziori E. & Dienes Z. (2015). Facial beauty affects implicit and explicit learning of men and women differently. *Frontiers in Psychology*, **6**, Article 1124.

Cite this article: Schulz, J. & Wonnacott, E. (2024). Pragmatic competence and pragmatic tolerance in foreign language acquisition—revisiting the case of scalar implicatures. *Applied Psycholinguistics* **45**, 717–744. https://doi.org/10.1017/S0142716424000274

https://doi.org/10.1017/S0142716424000274 Published online by Cambridge University Press