DEPOSIT AND CONSULTATION OF DISSERTATION

One copy of your dissertation will be deposited in ORA (Oxford University Research Archive), where it is intended to be freely available online. In order to facilitate this, you are requested to complete and sign the form below.

Please use block capitals

Surname SCHULZ
First names (in full) JOHANNES MAXIMILIAN
Faculty board EDUCATION
Degree name and pathway MSc APPLIED LINGUISTICS AND SECOND LANGUAGE ACQUISITION
Title of dissertation PRAGMATIC COMEPTENCE AND PRAGMATIC TOLERANCE IN FOREIGN LANGUAGE ACQUISITION – REVISITING THE CASE OF SCALAR IMPLICATURES N.B. The title stated here must be precisely the same as that stated on the title page of the thesis submitted. **A candidate wishing to amend the title previously approved by the faculty must apply to the faculty board for permission to do so.**
Supervisor ELIZABETH WONNACOTT
Subject keywords *Enter your own keywords or phrases to describe your work. This information helps us describe your work in ORA* SECOND LANGUAGE ACQUISITION, PRAGMATICS, SCALAR IMPLICATURES, PRAGMATIC COMPETENCE, PRAGMATIC TOLERANCE
Research methods used *This information helps us describe your work on SOLO for future students e.g. quantitative, interviews, vocabulary test, systematic review, etc.* QUANTITATIVE, GORILLA ONLINE PLATFORM

Declaration by the candidate as author of the dissertation

1. I understand that I am the owner of this dissertation and that the copyright rests with me unless I specifically transfer it to another person.
2. I understand that the Department requires that I shall deposit one copy of my dissertation in the Oxford University Research Archive ('ORA') where it shall be freely available online for use in accordance with ORA's Terms and Conditions of Use [https://ora.ox.ac.uk/terms_of_use].
3. I understand that this dissertation should not contain material that can be used to personally identify individuals or specific groups of individuals, and that such material should be removed before this dissertation is deposited in ORA.
4. I agree to be bound by the terms of the ORA Grant of Non-exclusive Licence [www.bodleian.ox.ac.uk/ora/deposit-in-ora/deposit-licence] and I warrant that to the best of my knowledge, making my thesis available on the internet will not infringe copyright or any other rights of any other person or party, nor contain defamatory material.
5. I agree that my dissertation shall be available for download in ORA in accordance with paragraphs 2, 3 and 4 above.

Signed [an electronic signature is sufficient]:
Date: 15/10/21 SCHULZ

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

Johannes Maximilian Schulz
University of Oxford
Linacre College

Word Count: 19627

Thesis submitted in partial fulfilment of the requirements for the Degree of
Master of Science in Applied Linguistics and Second Language Acquisition

August 2021

Candidate Number: 1049391

Table of Contents

Acknowledgements

I would like to thank my supervisor Elizabeth Wonnacott and her colleague Eva Viviani for their helpful advice and their always prompt, straightforward and very constructive feedback throughout the process of writing up this thesis. A solid and reliable team!

I thank Ryde School, UK, for their recruitment support with British students. Also, I am grateful to my former high school English teacher in Germany. When I asked for participant recruitment support, she did not hesitate and recruited her students. Her excitement, interest and support all these years after I left school is all a student can wish for in a teacher.

A pandemic is no piece of cake – even in Oxford. I am grateful to my friends in Oxford, back home in Germany, and in the Netherlands. Especially, thank you, Mikaela, my favourite Canadian, for bothering me each day and for being a wonderful friend.

I am truly thankful to my dear Anki, my longest and most loyal comrade. You are so affectionate and smart, and you have always stood by me. You have irrepressible believe in me, and I can count on you. Thank you.

Lastly, I would like to thank my family for their support – above all my parents. Besides ever-loving advice and financial support, my parents have provided my brother and me with a home where love has always been unconditional. I feel treasured and I am deeply grateful for this.

Abstract

Understanding language involves making inferences. One type of inference are scalar implicatures, e.g. understanding that "some X" generally implicates "some but not all X". Recent research has looked at the derivation of scalar implicatures employing binary sentence judgement tasks investigating differences between native speakers and L2 learners in terms of how they accept *some* sentences with a weak scalar expression (e.g. "Timothy ate *some* of the pretzels") in contexts where the stronger scalar expression would hold true (i.e. where Timothy ate *all* of the pretzels). Results have been inconclusive, but in part, findings have been taken as evidence that L2 learners are less pragmatically competent than native speakers regarding their scalar implicature derivation abilities. Following evidence from L1 acquisition research, I propose that the binary judgement tasks in L2 research did not measure pragmatic competence, but pragmatic tolerance. Therefore, to measure participants' pragmatic competence (i.e. sensitivity to underinformativeness) more unambiguously, I introduce into the field of L2 research the use of *graded* judgement tasks, previously employed in L1 studies. I present data from an experiment with adult English L1 speakers (n=30) and German EFL learners (n=36) judging underinformative *some* sentences in both binary and quinary sentence judgement tasks. Results with the quinary measure data find no evidence that the German EFL learners and native speakers differ in their pragmatic competence in terms of sensitivity to underinformativeness. Moreover, data with a binary measure task of the type used in previous L2 research also found no evidence for between-group differences in pragmatic tolerance. The findings suggest a reinterpretation of previous L2 scalar implicature research and reinforce the utility of distinguishing 'Pragmatic Competence' and 'Pragmatic Tolerance' in L2 experimental pragmatics.

List of Abbreviations

CUREC	Central University Research Ethics Committee
DREC	Departmental Research Ethics Committee
EFL	English as a foreign language
ERP	Event related potential
CEFRL	Common European Framework of Reference for Languages Proficiency levels: A1-A2 (basic user), B1-B2 (independent user), C1-C2 (proficient user) (cf. Council of Europe, 2021)
CP	Cooperative Principle
L1	Native language
L2	Second language
SJT	Sentence Judgment Task

List of Figures

List of Tables

1 Introduction

In communication, people constantly convey more meaning with what they say than is strictly conveyed by the meaning of the words they *utter*, and usually this is understood by the interlocutor. Investigating this phenomenon is one of the core areas of pragmatics. Often, meaning that goes beyond an utterance's literal proposition is conveyed through implicatures. For example, consider (1) and (2) which typically evoke the scalar implicatures (3) and (4), respectively:

(1) Peter will bring a salad or a cake to the party.
(2) Some men are good at driving a car.
(3) Peter will bring either a salad or a cake to the party, *but not both*.
(4) *Some but not all* men are good at driving a car.

From a theoretical pragmatics perspective, the connectives *and/or* and the quantifiers *some/all* are ordered on lexical scales which are ranked by informativeness in ascending order: <or, and>; <some, all>. The use of the weaker one of two 'scalemates' (e.g. *some* in <some, all>) implies the negation of the stronger term, evoking a scalar implicature such as *some but not all*. Generally, the production of the weaker term generally gives rise to the interpretation that the stronger term is not true, essentially because if the stronger term were true, the speaker would have used it (Doyle, 1951; Horn, 2006).

In contrast, from a formal logic perspective, *and* entails *or* and *all* entails *some*. Therefore, from this formal perspective the weaker term does not automatically preclude the stronger one. Thus, in formal logics, the connective *or* in (1) and the quantifier *some* in (2) would evoke the logical interpretations in (5) and (6), respectively:

(5) Peter will bring a salad or a cake to the party, *and possibly both.*
(6) *Some and possibly all* men are good at driving a car.

In sum, there are two possible interpretations of weak scalar terms such as *some* and *or*: a pragmatic interpretation (*some but not all; X or Y, but not both*) and a logical interpretation (*some and possibly all; X or Y, and possibly both*).

Researchers in experimental pragmatics have investigated how and why individuals arrive at either the logical or the pragmatic interpretation. Essentially, they investigated the psychological mechanisms that underlie inferencing processes such as pragmatic enrichments of scalar terms (cf. Holtgraves et al., 2019). This experimental inquiry has yielded several theoretical accounts of how scalar implicatures are cognitively processed in terms of factors such as processing speed and processing order, for example the default model (Levinson, 2000) and the non-default model (Sperber & Wilson, 1986).

Recently, a new strand of research has begun to investigate the interpretation of scalar terms in the context of second language (L2) acquisition. Understanding pragmatic abilities in L2 is of particular relevance since pragmatics remains a challenging area for L2 learners (e.g. Belletti et al., 2007). In L2 contexts, three main questions have been of research interest: (a) Are L2 learners able to derive scalar implicatures? (b) Are there differences between L1 speakers and L2 learners regarding their ability to derive scalar implicatures? (c) Which scalar implicature processing theory can account for the findings? Some such experiments have used target sentences such as *Some elephants have trunks*, asking whether participants tend to make the pragmatic or logical interpretation and how this differs from L1 speakers. Findings have been inconclusive to this date; some reported that L2 learners give *more* pragmatic answers than L1 speakers (e.g. Lin, 2016; Slabakova, 2010), while others found no differences between L2 learners and native speakers (e.g. Dupuy et al., 2019; Snape & Hosoi, 2018). Most recently, Mazzaggio et al. (2021) reported that L2 learners gave *fewer* pragmatic answers than L1 controls. The authors of these studies attempt to account for their findings through theories of scalar implicature processing such as the default and non-default models. However, given the inconsistency of results, the status of L2 learners' ability to derive scalar implicatures is not yet determined.

The current study adds to the work investigating the derivation of scalar implicatures by L2 learners. One notable feature of the L2 studies to date is that they all measured interpretation of implicatures using binary sentence judgement tasks

(SJT): Participants were usually presented with an underinformative *some* statement such as *Some elephants have trunks* and had to either agree or disagree with the statement (based either on world knowledge or some visual display). Regularly, disagreement is taken as evidence that participants have derived the scalar implicature (i.e. corresponding to the pragmatic interpretation: *Some but not all elephants have trunks*), while agreement is taken as evidence that participants have not derived the scalar implicature (i.e. corresponding to the logical interpretation: *Some and possibly all elephants have trunks*). However, being an L2 learner myself, I had the intuition that some participants in the studies above might have derived the implicature (and thus arrived at the pragmatic interpretation) but nevertheless *agreed* with the target statements because they might have decided not to categorically reject the statement but consciously accept the (minor) pragmatic violation. After all, a sentence such as *Some elephants have trunks* is not blatantly wrong. The issue is that for such a participant the binary SJT measures an aspect of metalinguistic judgement – and not the ability to derive scalar implicatures per se.

Interestingly, a similar line of research with children has investigated the development of pragmatic competence in L1 acquisition (e.g. Katsos & Smith, 2010). In this context, researchers noted that, from a theoretical linguistics perspective, binary SJTs are an insufficient instrument to gather data that determines whether someone is able to derive a scalar implicature or not (Katsos & Bishop, 2011). In more general terms, Veenstra and Katsos (2018) have argued that binary SJTs measure pragmatic *tolerance*, as opposed to pragmatic *competence*. These are distinct pragmatic concepts because pragmatic competence draws on general cognitive abilities, whereas pragmatic tolerance draws on metalinguistic attitudes. Moreover, L1 acquisition research has demonstrated empirically that *graded* judgement tasks (i.e. with more than two decision options) can provide a more unambiguous instrument to determine participants' pragmatic competence.

Returning to previous L2 research regarding scalar implicatures, these findings from the literature on L1 pragmatic acquisition raise the question of whether the relevant L2 research actually investigated L2 learners' pragmatic *competence* as

opposed to their pragmatic *tolerance* and whether this has contributed to the inconsistency of findings.[1]

The current study investigates the methodological issues raised by Katsos and colleagues, extending this to the context of L2 acquisition research. An experiment was conducted with German EFL learners and an L1 English comparison group in which participants responded to scalar implicature statements using both a binary and a graded SJT. By doing so, the study aims to demonstrate the following: First, contrary to some previous L2 findings, the graded SJT should show that L2 learners are in fact as pragmatically competent as native speakers. Second, the difference between L1 and L2 speakers found in previous L2 studies will be replicated in the binary judgement task. If such a pattern is found, this will confirm that between-group differences should be accounted for in terms of differences in pragmatic *tolerance*, as opposed to pragmatic *competence*. In this respect, the current study may also offer potential explanations for the inconclusive findings in previous L2 research. The study aims to demonstrate that L2 learners, at least those with high proficiency levels, can be unambiguously considered pragmatically competent if researchers are aware of the limitations of their research instruments. Additionally, it adds the concept of pragmatic tolerance from L1 acquisition research to the scope of L2 scalar implicature research.

[1] I corresponded with Dr. Katsos, one of the key critics of the use of binary judgements in investigating child L1 pragmatics – he agreed with my intuition that binary judgements in L2 research might have provided an insufficient instrument to unambiguously investigate questions regarding pragmatic competence.

2 Literature Review

This experimental pragmatics study aims to explore the interpretation of scalar terms by L2 learners, as well as the impact of experimental methods on previous research results and previous theoretical conclusions. The present chapter begins with a basic introduction to pragmatic implicature theory. This is followed by a review of literature addressing different psychological accounts of implicature processing, as well as previous empirical findings concerning scalar implicature derivation in both L1 and L2 experimental pragmatics research. The chapter will then provide an overview of criticisms of experimental paradigms investigating scalar implicature interpretation and will introduce the Pragmatic Tolerance Hypothesis. Against this critical background and in the framework of pragmatic tolerance, I will suggest novel theoretical and experimental directions in L2 scalar implicature research.

2.1 Grice's Implicature Theory and Quantity Implicatures

Grice (lectures given in 1967, first published in 1975) introduced the theory of conversational implicatures. It accounts for the phenomenon that *what is meant* in a conversation often exceeds *what is literally said* (Breheny, 2019). Consider (1), a classic Gricean example (1975: 32; '+>' means 'conversationally implicates'):

(1) A: I am out of petrol.
B: There is a garage around the corner.
+> (1') The garage is open and it sells petrol.

Typically, a listener would understand B's utterance to mean that the garage is (a) open and (b) sells petrol, although A has not literally made these propositions (Cummins, 2019). This listener-centred interpretation process is guided by an assumption that the speaker follows an underlying communication principle, the so-called *cooperative principle* (CP): "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" (Grice, 1975: 45). The principle underlies rational and effective communication, it guides communicators, and it is mutually presupposed by the communicators during communication (Green, 1990). Additionally, Grice (1975) established four maxims which – when a rational speaker is being cooperative –

instantiate the CP (see Figure 1). They are neither a set of rules nor norms that we *consciously* stick to. Rather, they instantiate cooperative behaviour (Green 1990).

Fig. 1 Grice's (1975) four conversational maxims

I Quantity:

1. Make your contribution as informative as is required (for the current purposes of the exchange).
2. Do not make your contribution more informative than is required.

II Quality: Try to make your contribution one that is true:

1. Do not say what you believe to be false.
2. Do not say that for which you lack adequate evidence.

III Relation: Be relevant.

IV Manner: Be perspicuous.

1. Avoid obscurity of expression.
2. Avoid ambiguity.
3. Be brief (avoid unnecessary prolixity).
4. Be orderly.

As long as a speaker (unconsciously) adheres to these maxims, i.e. as long as s/he is being cooperative, the hearer does not notice them. Accordingly, example (1) would be considered a *relation implicature* (other types of conversational implicature include *quality implicatures*, *manner implicatures* and *quantity implicatures*) because (1') is the most relevant interpretation of B's utterance. The implicature derives because the speaker adhered to the maxim of relation: the utterance is relevant in this specific context. Although there are other ways to generate implicatures (e.g. through blatantly flouting maxims; cf. Cummins, 2019; Grice, 1975; Kallia, 2007), adherence to maxims is the only relevant mode in the current context because adherence to maxims is the typical trigger of the quantity implicatures which are the topic of this thesis.

Quantity implicatures typically arise in situations where a speaker uses a less informative expression although s/he could have used a more informative one, but chose not to. Consider example (2):

(2) Timothy ate some of the pretzels.

+> (2') Timothy did not eat all of the pretzels.

Under the presumption that the speaker (A) is a cooperative communicator, i.e. A is sane and not blatantly trying to trick the hearer (B), B would usually infer that Timothy did not eat all of the pretzels, although A did not say so (Liedtke, 2016). The fact that A has used the less informative term *some* instead of the more informative *all*, leads B to infer that A thinks that an utterance with the more informative term *all* would be wrong. This is because A is assumed to adhere to the Maxime of Quantity (viz. to be a cooperative interlocutor), meaning that s/he would have used the more informative term *all* if s/he believed it to be true. Guided by A's choice of the term *some*, B thus either (a) believes that A does not know whether Timothy ate all of the pretzel or (b) infers that Timothy *did not eat all of the pretzels*. Generally, along a continuous scale of quantification terms such as <none, some, many, all>, the use of weaker terms leads to the inference that neither of the stronger terms would hold true in that particular situation (e.g. in a situation where Timothy did not eat all of the pretzels). This scale-based deduction process is referred to as *scalar implicature* (Horn, 1972 – as cited in Horn, 2006, p. 8). Conversely, this means that if A states (2) and Timothy has in fact eaten *all* of the pretzels, utterance (2) violates the maxim of quantity and can be said to be 'not maximally informative'/'underinformative' (Slabakova, 2010). Other striking examples of underinformative utterances are those where there is an incongruence of scale-term and context based on encyclopaedic knowledge:

(3) Some elephants have trunks.

+> (3') Some but not all elephants have trunks.

Under the assumption that the person who uttered (3) is sane and cooperative, (3) seems 'odd': The inference that *some but not all elephants have trunks* contradicts our knowledge that each elephant usually possesses a trunk – making the utterance underinformative. Note that the inferencing processes in (2') and (3') initially require the listener to realize that a more informative term (e.g. *all*) could have been used in the first place. This is referred to as *sensitivity to informativeness* (Davies & Katsos, 2010). Without this precondition, an implicature derivation process cannot start.

2.1.1 Horn scales

Although intuition might tell us that the utterances (2) and (3) are somewhat odd (in the respective contexts), they are not blatantly wrong (Magri, 2009; 2011) but instead *pragmatically infelicitous*. From a logical perspective, they are not incorrect since one could infer, for example, that *Some and possibly all elephants have trunks*. Essentially, the utterances could be interpreted both pragmatically (*some but not all*) or logically (*some and possibly all*). This problem is systematically described in terms of so-called *Horn scales*. Horn (1972) maintained that particular kinds of stronger expressions logically entail respective weaker expressions and categorized them on continuous quantification scales. Regarding the scale <some, all>, *all* logically entails *some*. Other scales include items such as modals <can, must> or adjectives <warm, hot>. Against the background of such scales, the logical interpretation of some (*some and possibly all*) is considered lower-bound while the pragmatic interpretation (*some but not all*) is considered upper-bound. Therefore, based on Horn scales, the scalar implicature in (2') contains the *upper-bound* interpretation of *some*.

The theories so far stem from theoretical linguistics. From a psychological perspective, researchers are interested in which interpretations and inferences speakers actually make when encountering utterances that include Horn scale items such as *some*, and also why they make them. These inquiries are investigated in experimental pragmatics. Experiments typically involve situations where a speaker describes a situation with an underinformative *some* utterance, although *all* would be the most informative term. Experimenters then test whether (and how fast) participants draw an upper-bound (pragmatic) or lower-bound (logical) inference. These data are used to make assumptions about the cognitive scalar implicature derivation processes and to inform relevant pragmatic theory. In the following section, I will describe the two main implicature processing theories researchers have developed from research in experimental pragmatics which is typically based on the theoretical linguistic theories discussed above.[2]

[2] The experimental literature sometimes uses the terms *inference* and *implicature* interchangeably, therefore, it is important to make some clarifying theoretical remarks regarding their precise and distinctive meanings. On the one hand, implicatures are an aspect of speaker meaning which goes beyond the literally uttered words and they are restricted to the speaker's part of the communicative act

2.2 Two different accounts of implicature processing

Grice's (1975) theoretical work on implicatures and Horn's (1972) development of scales were not focused on cognitive processes per se. In this section, I will turn to two theories that build on those earlier theories and incorporate more cognitive accounts of the processing of scalar implicatures: the default and the non-default model. In the past, experimental pragmatics research looking at scalar implicature processing often set out to test these two models.

2.2.1 Default Model

The default model (Levinson, 2000) is based on a Neo-Gricean pragmatic framework (Horn 1972; Gazdar, 1979) and proposes that pragmatically enriched meanings are derived by default irrespective of the context (see Figure 2). That is, as soon as a listener receives a scalar term such as *some*, the implicature is automatically derived, i.e. even before s/he has finished reading/listening to the sentence, s/he has arrived at the pragmatic meaning of the scalar expression (*some but not all*). In terms of Horn's scales, receiving a weaker term of a Horn scale immediately triggers the negation of stronger terms of the same scale (Cummins, 2019). In the default model, this immediate derivation process is based on Levinson's (2000) Q-heuristic (*What isn't said, isn't*), a Neo-Gricean derivative of Grice's maxim of quantity.

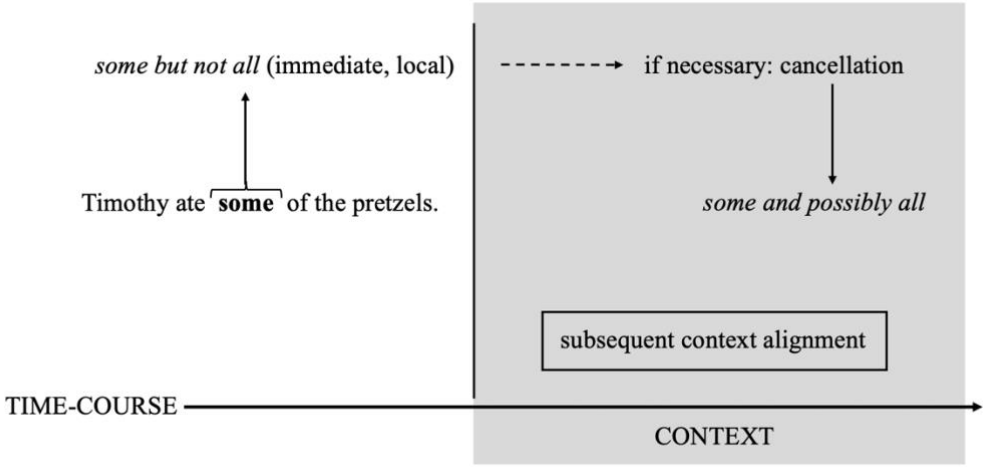
The Q-heuristic prevents speakers from using an informationally weaker term in a situation where an informationally stronger term would more appropriately reflect their knowledge, and thus in example (2), the Q-heuristic immediately leads the listener to

(Horn, 2006; Haugh, 2013; Skordos & Papafragou, 2016). On the other hand, inferences are the cognitive processes on the part of the hearer to deduce meaning beyond what is literally uttered (Haugh, 2013). Via an inferencing process (e.g. by considering context propositions and potential speaker intentions), the hearer arrives at the so-called *implicatum* (a term introduced by Grice) which is the product of his/her inferencing process (cf. Grice, 1975; Haugh, 2013). Essentially, a speaker (S) implicates and a hearer (H) infers (Haugh, 2015). In successful communication, the implicature (on the part of S) matches with the inferred implicatum (on the part of H). Against this conceptual background, the current study technically examines the *implicatum* which L2 learners infer from a target utterance which carries an implicature. For instance, Mazzaggio et al. (2021: 4) refer to the 'computation of scalar implicatures' (but also Dupuy et al., 2019; Slabakova, 2010), however, technically, this is imprecise, as what is computed via the hearer's inferencing process is an *implicatum* (not the implicature), which, ideally, matches the *implicature* a speaker intended to make (Haugh, 2013). As the current study is not concerned with theoretical pragmatic notions and to prevent confusion, I will also use the term *implicature* when referring to the product of a listener's utterance-interpretation, that is, the pragmatically enriched meaning.

infer: 'Because A has not used the more informative term *all*, it is not *all*' – without awaiting the sentence's end. Importantly, in the default model the pragmatic interpretation – or implicature (i.e. the upper-bound pragmatic meaning of a weak scalar term) – is triggered directly when the listener hears the lexical triggers (e.g. *some*) during sentence processing (Cummins, 2019), a process called *local enrichment*.

If this pragmatic interpretation clashes with the context conditions, it can be cancelled to return to the logical meaning of the scalar term (i.e. lower-bound). However, this cancellation will occur at a subsequent processing stage (*subsequent context alignment* – see Figure 2) which increases the overall processing time (Cummins, 2019; Mazzaggio et al., 2021).

Fig. 2 Default model illustration of scalar implicature processing

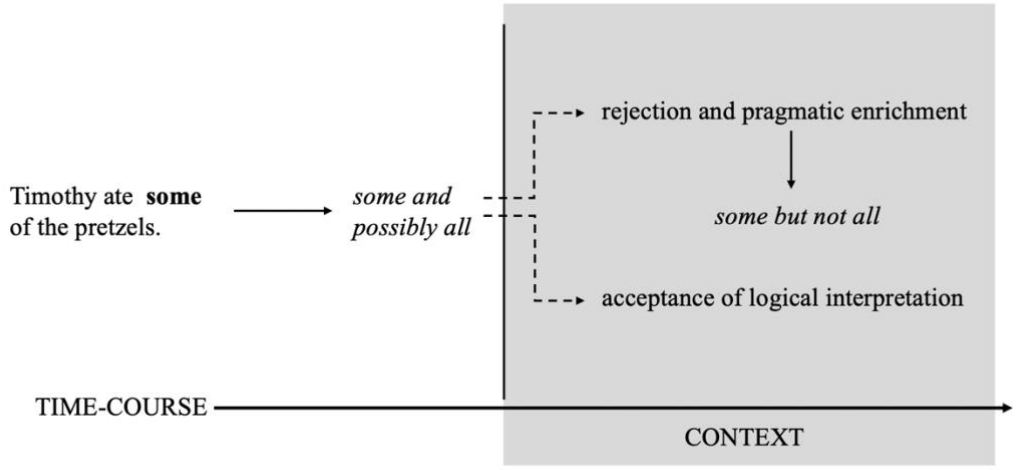


2.2.2 Non-default Model

The non-default model is an alternative processing account based on Relevance Theory of pragmatics rather than on a Neo-Gricean approach. Relevance Theory (Carston, 1998; Sperber & Wilson, 1986; 2002) reduces Grice's Implicature Theory simply to: 'Be relevant'. This means that in a context where both *some* and *all* are true, *all* is more relevant than *some*, thus, *all* is to be used. On the other hand, if the context indicates that only *some* is true, *all* is irrelevant and simultaneously negated by virtue of the speaker's use of *some* (Geurts, 2010). In contrast to the default model, the non-default model proposes no default pragmatic enrichment of lexical triggers such as *some* along the lexical scale <some, all>: there is no *automatic* triggering of pragmatic

enrichment as in the default theory. On the non-default account, recipients first process the entire sentence and arrive at the logical meaning of the scalar term (i.e. lower-bound). Subsequently, the context conditions to warrant an implicature have to be satisfied to arrive at the pragmatically enriched meaning of the scalar term (i.e. upper-bound) (Cummins, 2019). Essentially, after receiving the sentence, a hearer (implicitly) asks whether the logical interpretation is relevant or whether pragmatic enrichment would be more relevant in the current context (see Figure 3). Importantly, this is a *global enrichment* process, as the comparison of the sentence's proposition to the context happens *after* the entire sentence has been processed (Chemla & Singh, 2014b). Therefore, the non-default model is also referred to as *contextualist* approach.

Fig. 3 Non-default (i.e. contextualist) model illustration of scalar implicature processing



2.2.3 Default & Non-default in L1 experimental research

The two theories make distinctive predictions regarding the derivation of scalar implicatures as both the default-characteristic and the non-default-characteristic are considered to impact processing speed (Breheny, 2019; Chemla & Singh, 2014a). On the one hand, under the default view the pragmatic interpretation is derived by default, thus, deriving this interpretation requires fewer processing steps (provided it does not clash with context), and thus, is less time consuming than deriving the logical interpretation of *some*. On the other hand, under the non-default view the pragmatic interpretation is not derived by default and thus requires more processing steps, and is more time consuming than deriving the logical interpretation of *some* (Cummins, 2019; Dupuy et al., 2019).

The approaches' predictions have been investigated with native speakers in numerous experiments, for example, through truth value judgements (Bott & Noveck, 2004; Bott et al., 2012), self-paced readings (Breheny et al., 2006), reaction time and eye-tracking measurements (Degen & Tanenhaus, 2011, 2015, 2016; Huang & Snedeker, 2009), neurolinguistic ERP (event-related potential) studies (Noveck & Posada, 2003) and visual-world studies (Grodner et al., 2010; Storto & Tanenhaus, 2005). Generally, results support the non-default view and contradict the default-model as evidence has suggested that scalar implicatures are derived neither automatically nor context-independently, but that the derivation is delayed and simultaneously context-dependent (Bott & Noveck, 2004; Bott et al., 2012; Huang & Snedeker, 2009). For example, in Noveck and Posada's (2003) ERP study, 19 French adults were presented with categorical statements such as *Some elephants have trunks* and had to agree or disagree. Reaction times indicated that disagreement with an underinformative *some* statement (i.e. pragmatic interpretation *some but not all*) took significantly longer than agreement (i.e. logical interpretation *some and possibly all*). Regarding ERPs, neurolinguistic research typically expects that semantic anomalies in stimulus sentences evoke central parietal negative-going peaks about 400ms after an anomaly's appearance (so-called *N400*: Kutas & Hillyard, 1980a; 1980b). For example, in the context of a stimulus sentence such as *John buttered his bread with socks* it is expected that the ERP curve shows a negative-going peak 400ms after the appearance of the inappropriate word *socks*. Therefore, Noveck and Posada (2003) argued that if scalar implicatures are automatically triggered by the lexical trigger *some*, they should find an N400 relative to the trigger word *some* in the participants' brain activities because *some* would trigger a meaning that goes beyond its semantic content, i.e. it is considered a semantic anomaly. However, results showed no particular N400 reaction for both logical and pragmatic answers to underinformative sentences. Therefore, the authors conclude that scalar implicatures are not derived automatically at the word *some* but *after* sentence processing, a conclusion which also fits with the increased reaction times for pragmatic answers. This supports the non-default model (i.e. scalar implicature processing is delayed).

Using similar input (e.g. *Some elephants are mammals*), Tomlinson et al. (2013) employed a mouse-tracking technique in a sentence verification paradigm

(*agree/disagree*) with 40 English native speakers. The study showed that when participants disagreed with underinformative sentences (i.e. pragmatic response), they moved their mouse first to the agree-button before moving to the disagree-button. In contrast, when they gave logical responses, their mouse moved directly to the agree-button. The authors take this observation as evidence in favour of the non-default model as this model posits that a pragmatic interpretation of *some* (*some but not all*) presupposes the cancellation of the logical interpretation (*some and possibly all*), as demonstrated by the mouse-movements.

2.3 Scalar implicature derivation in L2 research

We have seen that adult L1 studies seem to support the non-default view of scalar implicature processing. In contrast, some recent research with adult L2 learners seems to support the default view, although the findings here are inconsistent (e.g. Dupuy et al., 2019; Lin, 2016; Slabakova, 2010; Snape & Hosoi, 2019). Similar to L1 studies, L2 studies have used the phenomenon of 'underinformativeness' to examine participants' scalar implicature derivations and have interpreted the results in light of the default and non-default models. The aim has been to determine how L2 learners' pragmatic abilities, in particular their ability to derive scalar implicatures, compares with native speakers. The incentive for such research is the fact that pragmatics remains a persistent challenge to L2 learners even at advanced stages. For example, L2 studies have shown that even near-native L2 learners struggle with discourse-pragmatics constraints of null-subject languages during immediate discourse, which is manifested in overuses of overt subjects compared to native speakers (e.g. Belletti et al., 2007). These observations are notwithstanding the fact that L2 participants usually master the null-subject-parameter syntactically (Belletti et al., 2007; Gürel, 2006; Lozano, 2006) – i.e. essentially, proficient L2 learners know how to use null-subjects in an L2, however, they misuse the parameter in pragmatically inappropriate ways. Generally, L2 research suggests that adult L2 learners have fewer processing resources available for language processing than native speakers (Clashen & Felser, 2006). This cognitive L2 'disadvantage' leads to processing disadvantages in immediate communication (cf. above *null-subjects*) and this may have implications for processing of scalar implicatures: on the one hand, if L2 learners give more pragmatic answers to

underinformative *some* sentences than native speakers, this is considered evidence in favour of Levinson's (2000) default model, because cancelling automatically derived implicatures requires additional effortful processing steps and this cost will be more marked for L2 learners who have fewer resources available for this process. On the other hand, if L2 learners give fewer pragmatic answers to underinformative *some* sentences than native speakers, this is considered evidence in favour of the non-default model because in this account it is deriving scalar implicatures which is more effortful since it happens after sentence processing and requires the cancellation of the logical interpretation through context alignment. In the following, I shall give an overview of main L2 research findings and discuss potential reasons for their inconsistency.

2.3.1 Previous L2 findings

Using an SJT, Slabakova (2010) compared judgements of underinformative *some* sentences by adult English native speakers (n=23) and adult Korean L2 learners of English (n=50). The L2 group consisted of thirty advanced learners of English and twenty intermediate learners of English. In addition, a Korean native speaker group (n=30) were tested in Korean to control for culture-specific interpretation bias. Participants had to either agree or disagree with eight optimally true sentences (*All elephants have trunks*), eight infelicitous (i.e. underinformative) *some* sentences (*Some elephants have trunks*), eight felicitous *some* sentences (*Some books have color pictures*), eight optimally false sentences (*All books have color pictures*) and eight absurd fillers (e.g. *All/some garages sing*). All statements were categorical sentences based on encyclopaedic knowledge and no context was provided. The underinformative *some* sentences were the target items. In a second experiment, the author tested new sets of adult English native speakers (n=20), Korean learners of English (n=36 advanced, n=20 intermediate) and Korean native speakers (n=35). The task remained the same, however, visual context in the form of short picture-stories was provided. Again, participants had to judge the same types of sentences, including underinformative *some* items, for example, *I've eaten some of the candies* (in a context where the protagonist in the picture-story had eaten all of the candies). While no effects of L2 proficiency were detected, L2 learners chose the pragmatic answer option (i.e. disagree to underinformative *some*) significantly more often than both the English

monolingual group and the Korean control group tested in their L1, in both experiments. As an example: in the first experiment, L2 learners disagreed with sentences such as *Some elephants have trunks* more often than both other groups. In the second experiment, L2 learners disagreed with sentences such as *I've eaten some of the candies* (although the character had actually eaten all of the candies) more often than the other two groups. As discussed above, disagreement with those underinformative *some* statements is considered the pragmatic (i.e. upper-bound *some but not all*) interpretation. Since L2 learners are considered to have higher cognitive burdens compared to L1 speakers, Slabakova (2010) took the fact that L2 learners gave more pragmatic answers than the comparison L1 groups as evidence in support of the default model which postulates that pragmatic interpretations are automatic and 'easier' default interpretations.

Slabakova's (2010) findings are confirmed by Lin (2016) who tested Mandarin L2 English learners' (proficiency: B1-B2) judgements of underinformative *some* sentences in a binary SJT, again finding that L2 learners gave significantly more pragmatic answers than native speaker controls. Following Slabakova's (2010) line of argument, Lin (2016) interprets these results as evidence in support of the default model.

Dupuy et al. (2019) also employed binary SJTs to test pragmatic bias in L2 learners compared to native speakers. Their three participant groups – native speakers of French with no L2 knowledge (n=30), native speakers of French learning English as L2 (n=30, proficiency: B2) and native speakers of French learning Spanish as L2 (n=30, proficiency: B2) – had to give 'Yes'/'No' answers to underinformative *some* statements (e.g. *The boy has hidden some cars*) relative to visual contexts. The monolinguals were tested in French only, the L2 groups were tested in both French and in either English or Spanish (with the presentation-order of the two languages (i.e. mother tongue first and L2 second, or vice versa) counterbalanced). Importantly, pre-experimental instructions informed the L2 participants that they would be tested in both their L1 and L2. Results indicated that both L2 learner groups gave more pragmatic answers in both their L1 (French) and L2 (English or Spanish) compared to the monolingual French group. Although the results in the L2 seem to confirm Slabakova's (2010) results, note that the L2 groups also gave the same proportions of pragmatic answers in their L1.

This contradicts Slabakova's (2010) conclusion in support of the default model: according to this model we would expect a difference in pragmatic answering depending on the language participants are tested in because of processing difficulties in L2. Further investigating this issue, Dupuy et al. (2019) conducted another experiment with similar tasks and stimuli, however, now their participants (French learners of English, n=46) were tested *either* in their L1 or L2. This between-subject experiment yielded no difference between the two language groups in the proportions of pragmatic answers, thus demonstrating no pragmatic bias in L2 learners, in contrast to both their previous experiment and Slabakova (2010). Dupuy et al. (2019) argue that their findings support neither the default model nor non-default model because both theories predict that L2 speakers either give more pragmatic answers (default) or more logical answers (non-default) than native speakers. Instead, the authors explain their findings in terms of metalinguistic awareness. Regarding the results from their first experiment (within-subjects), they assert that the intra-experimental language switch between L1 and L2 made participants more aware of pragmatic anomalies because it temporarily might have increased their concentration and metalinguistic awareness, resulting in pragmatically biased answer behaviours. Regarding their second experiment (between-subjects), they argue that learning an L2 increases one's metalinguistic awareness in general and, therefore, compensates for the lack of proficiency compared to native speakers, resulting in equal answer behaviours. They do not provide reasons as to why they got a different result in their second experiment compared to Slabakova (2010) and Lin (2016).

Similarly, Snape & Hosoi (2018) compared Japanese L2 learners of English (proficiency: B2) with English monolinguals regarding their judgement of underinformative *some* sentences. Their participants had to give a 'Yes'/'No' response to target underinformative *some* items (*Are some of the strawberries in the red circle*?) relative to visual contexts. In line with Dupuy et al.'s (2019) second experiment, the authors found no significant differences in answer behaviour between the language groups. Nonetheless, because their descriptive data suggested that the L2 group answered more pragmatically than the L1 group, they argue in support of the default model in that pragmatic interpretations are less costly and the default for L2 learners.

Although this conclusion conforms to Slabakova (2010), it is based on descriptive data and should be treated cautiously.

Most recently, Mazzaggio et al. (2021) compared the frequency of pragmatic answers (i.e. rejection of underinformative *some* statements) of Italian monolinguals to Italian L2 learners of English and Spanish in an aural SJT. Participants heard categorical statements (e.g. *Some elephants have trunks*) without visual context and agreed ('Yes') or disagreed ('No') with the utterances within a three-second time limit. The authors argue that both the aural character and the time limit add to the participants' cognitive load which would help to interpret the results more unambiguously regarding the two scalar implicature processing theories. Essentially, if the pragmatic interpretations of underinformative *some* statements are the default interpretations, deliberately increasing the cognitive load should result in extremely high rejection rates of underinformative items for L2 learners compared to L1 speakers. The same logic applies vice versa for the non-default model. Interestingly, contrary to prior studies, Mazzaggio et al.'s (2021) results indicated that L2 learners gave *fewer* pragmatic answers compared to native speakers. According to the authors, this finding supports the non-default model because it predicts that logical interpretations of underinformative *some* items (*some and possibly all*) – which lead to agreement with such an item – are the default interpretation, and that pragmatic enrichment after sentence processing is cognitively more costly. They argue that their results contradict Slabakova's (2010) findings because Slabakova's experimental L2 groups had lived in the target country for a considerable amount of time prior to the experiment. Accordingly, immersion might have been advantageous regarding their implicature processing abilities. That is, the authors argue that immersed bilinguals have "general metacognitive advantage[s]" (Mazzaggio et al., 2021, 31; cf. Adesope et al., 2010). Therefore, Mazzaggio et al. (2021) argue that their *non*-immersion participants had a sufficiently higher cognitive burden than Slabakova's (2010) participants which led their participants to make more logical interpretations (as per the non-default theory). This immersion-related conclusion is unexpected, because although research has shown that immersion increases L2 learners' pragmatic abilities (e.g. Bouton, 1992), it remains unclear why L2 learners would 'outscore' native speaker controls and derive *more* pragmatic interpretations in Slabakova's (2010) experiment. In fact, no

differences should be found (cf. Dupuy et al., 2019; Snape & Hosoi, 2018). Note also that explaining the difference between the findings in Mazzaggio et al. (2021) and Slabakova (2010) in terms of immersion-factors does not fit with the fact that, as noted above, Lin (2016) found a very similar pattern of results to Slabakova (2010), yet like Mazzaggio et al. (2021) used non-immersion participants.

Additionally, it cannot be ruled out that the introduction of additional cognitive loads (aural presentation and time limits) influenced L2 participants' behaviour in Mazzaggio et al.'s (2021) study. However, although it seems plausible to expect that participants provide more logical answers when the cognitive burden is increased, the 'increased cognitive burden factor' does not explain why Slabakova (2010) *without* increased cognitive burden reports *reverse* results – instead of only weaker results.

Another general factor which might have influenced findings across all studies is the provision of visual context. Some studies employed categorical stimulus sentences without visual context while others provided visual context. Categorial sentences such as *Some elephants have trunks* draw on participants' encyclopaedic knowledge, they do not require participants to evaluate immediate visual context. These sentences, however, have evoked criticism. For example, Guasti et al. (2005) argue that, in theory, if there is no visual context, participants could quickly conjure up alternative contexts in their mind where, for example, *not all elephants have trunks* (injured elephants in zoos perhaps). In such imagined contexts, *disagreement*/*agreement* describe mental processes which are not under investigation (e.g. creativity). Perhaps the provision of context influences participants' interpretation processes, thus, impacting findings. Interestingly, the studies that provided context found no significant differences in answer behaviour between L1 and L2 groups (Dupuy et al., 2019; Snape & Hosoi, 2018), while studies without visual context found significant differences (Lin, 2016; Mazzaggio et al., 2021; Slabakova, 2010 (experiment 1)). However, note that Slabakova's (2010) second experiment provided visual context and still found significant differences between L1 and L2 groups. Therefore, the notion that findings are inconsistent because of varying context conditions might only be one of several influential factors at play.

Overall, existing L2 studies reported inconsistent findings: some find *more* pragmatic answers in L2 groups compared to L1 groups, some *no* language differences,

and some *fewer* pragmatic answers in L2 groups compared to L1 groups. Possibly, several factors such as participants' metalinguistic awareness, influence of immersion, increased cognitive load or context provision influenced results, but exactly how these lead to the observed patterns of differences is unclear based on the research available. However, one factor that *all* studies described thus far have in common is that they employ *binary* SJTs where participants categorically agree or disagree with underinformative *some* statements (in encyclopaedic or visual contexts). Based on these binary data, the researchers derive insights into scalar implicature derivation processes, for example, Mazzaggio et al. (2021: 30; my italics) assert that "the decrease in pragmatic interpretations of underinformative sentences *can be taken as evidence that deriving such pragmatic interpretations is costly and non-automatic*" and make claims regarding L2 learners' pragmatic competence, such as "[some] learners are able to calculate the implicature" (Snape & Hosoi, 2018: 186).

In contrast to the relevant L2 studies to this date, I build the current thesis, including my experimental inquiry, on the argument that these L2 studies employed an insufficient experimental method which is insufficient to make such claims about L2 pragmatic competence and scalar implicature processing theories. I argue that neither pragmatic competence nor implicature theories can be informed by data generated with binary SJTs, and that the inconsistency of findings, in part, stems from this methodological insufficiency. My argument is based on a similar line of research with L1 children which identified methodological flaws in the field of scalar implicature processing research as a whole (e.g. Katsos & Bishop, 2011; Katsos & Smith, 2010; Pipijn & Schaeken, 2012) and sparked fundamental experimental paradigm criticism (e.g. Veenstra & Katsos, 2018; Waldon & Degen, 2020).

2.4 Binary Sentence Judgement Tasks – Paradigm Criticism

In recent years, researchers have critiqued the use of SJTs when examining monolingual children's and adults' pragmatic abilities in experimental pragmatics research (e.g. Veenstra & Katsos, 2018). Regarding scalar implicatures, where the quantifier *some* can be interpreted logically (*some and possibly all*) and pragmatically (*some but not all*), Schmitt and Miller (2010) argue that agreement with underinformative *some* statements such as *Some elephants have trunks* does *not*

necessitate that participants have not generated the implicature: They may have generated the pragmatically infelicitous scalar implicature *Some but not all elephants have trunks*, but at the same time they are willing to tolerate and accept this pragmatic interpretation because the pragmatic violation is not considered grave enough to reject the sentence. Essentially, perhaps someone has the pragmatic competence to derive scalar implicatures, but nonetheless *chooses* to accept a statement based on a metalinguistic attitude that the violation is acceptable and warrants no rejection. Thus, experimental outcomes of binary SJTs do not reveal participants' pragmatic competence (i.e. whether s/he is able to derive scalar implicatures). Instead, they reveal participants' *tolerance* for pragmatically infelicitous interpretations. Table 1 illustrates possible relationships between reasoning processes and behaviours given different types of judgement task (note: *sensitivity to underinformativeness* and *graded SJTs* are discussed below). Importantly, these processes in part rely on metalinguistic attitudes and not on pragmatic abilities per se (Katsos & Bishop, 2011).

Table 1. Participants' reasoning processes and reactional behaviour in binary and graded SJTs. Information adapted from Veenstra and Katsos (2018).

Input	**Some elephants have trunks.**					
Reasoning processes	Participant does not even realize that a more informative quantifier (here: *all*) could be used and accepts the statement.*	Participant is not sensitive to the underinformativeness of the statement and does not generate the scalar implicature but interprets the statement logically, i.e. *Some and possibly all elephants have trunks* and accepts the statement.	Participant is sensitive to the underinformativeness of the statement (but does not generate an implicature), that is, s/he notes that a more informative quantifier could be used but chooses to tolerate this pragmatic violation based on his/her metalinguistic attitude that the violation is **not** grave enough to warrant a categorical rejection.	Participant is sensitive to the underinformativeness of the statement, generates the scalar implicature *Some but not all elephants have trunks* and notes the interpretation's pragmatic infelicity. However, the participant accepts the statement based on his/her metalinguistic attitude that this violation is **not** grave enough to warrant a categorical rejection.	Participant is sensitive to the underinformativeness of the statement, notes that a more informative quantifier could be used and based on his/her metalinguistic attitudes choses to categorically reject the statement before generating an implicature.	Participant is sensitive to the underinformativeness of the statement, generates the scalar implicature *Some but not all elephants have trunks* and notes the interpretation's pragmatic infelicity. Based on his/her metalinguistic attitudes the participant chooses that this violation is grave enough to warrant a categorical rejection.
Is scalar implicature derived?	No	No	No	Yes	No	Yes
*Reaction in **binary** task*	**Agree** → participant = not pragmatically **competent** (reasoning in previous studies)				**Disagree** → participant = pragmatically **competent** (reasoning in previous studies)	
Sensitive to underinformativeness?	No	No	Yes	Yes	Yes	Yes
*Reaction in **graded** task*	**Agree** → participant = pragmatically **oblivious/incompetent** (reasoning in current study)		**Intermediate** → participant = pragmatically **competent** and **tolerant** (reasoning in current study)		**Disagree** → participant = pragmatically **competent** yet **intolerant** (reasoning in current study)	

*While it is a theoretical possibility to presume that participants *genuinely* do not realize that one could also say *All elephants have trunks*, it is unlikely that someone without cognitive impairments behaves like this.

As demonstrated in Table 1, conclusions as to whether participants are able to derive scalar implicatures or not are impossible based on data collected in binary SJTs (see critical row in Table 1: "Is scalar implicature derived?"). The outcomes 'Agree' and 'Disagree' cannot unambiguously be linked to participants' abilities to derive implicatures.

Such methodological drawbacks of binary SJTs were first noticed by researchers exploring children's scalar implicature derivation abilities. A body of research in this field has appeared to demonstrate that children acquire the ability to interpret underinformative sentences *pragmatically* only after the age of seven, i.e. after they have acquired the ability to interpret the same sentences *logically* (e.g. Barner et al., 2011; Guasti et al., 2005; Noveck, 2001; Papafragou and Musolino, 2003), suggesting that pragmatic abilities regarding scalar implicatures do not develop until around the age of seven (Katsos & Bishop, 2011). Evidence comes from Papafragou and Musolino (2003), for example, who employed a binary SJT to compare Greek children's (mean age: 5;3) acceptance rates of underinformative *some* sentences such as *Some of the horses jumped over the fence* (visual context provided) to rates of Greek adult controls. The fact that the child group accepted target items significantly more often than the adult controls is considered evidence for young children's failure to derive scalar implicatures. Although such child research has not explained results in terms of the default/non default theories, the conclusion drawn from child research is similar to adult L2 research: both L1 children and L2 adults tend to fail to derive scalar implicatures due to cognitive disadvantages (Papafragou & Musolino, 2003). However, regarding child research, Veenstra and Katsos (2018) argue that for children this result is surprising since although younger children (1.5 – 2 years of age) show difficulties in some areas of pragmatics (e.g. regarding requests for clarification when encountering problems of understanding; cf. Ninio & Snow, 1996), in other areas these abilities are reasonably well-developed and well-documented even in two-year-olds. For example, Southgate et al. (2010) found that 17-months-old children used presumed mental states (about their interlocutors) to infer intended meaning, while Grassmann et al. (2009) report that two-year-olds draw on a conversation's common ground (a pragmatic principle of shared information) to include/exclude novel words in their vocabulary. Based on such demonstrations of children's pragmatic abilities, Veenstra and Katsos

(2018) argue that children's apparent inability to interpret underinformative *some* sentences pragmatically could be due to an experimental artifact, namely, the insufficient use of binary SJT tasks. Like Schmitt and Miller (2010), Veenstra and Katsos (2018) assert that accepting an underinformative statement does not unambiguously indicate that a child does not have the ability to generate scalar implicatures (see Table 1).

2.4.1 Criticism – Empirical evidence

Katsos and Bishop (2011) provided first evidence that using *non-binary* SJTs can reveal different patterns of pragmatic behaviour. They conducted two L1 experiments with adults and children using both a binary and ternary SJT. The ternary SJT offered an additional 'intermediate' decision possibility. In both experiments, in playful scenarios appropriate for children, participants for example saw a hungry giraffe which ate all of the pears from a tree but not the apples. After the scene, the experimenter asked an animated figure (Mr. Caveman) *What did the giraffe eat*? In the target underinformative *some* condition, Mr. Caveman would answer *The giraffe ate some of the pears*. In experiment 1 (children n=20, mean age: 5;6; adults n = 20), the participants' task was to reward Mr. Caveman for his answer by evaluating his statements with *that's right* or *that's wrong* (binary response). In experiment 2 (children n=18, mean age: 5;8; adults n=10), the participants had to reward Mr. Caveman for his answer by giving him one of three strawberries of different sizes (small, medium, large – representing a ternary SJT, where giving the middle-sized strawberry allowed an option for giving an intermediate level of endorsement). Results from experiment 1 confirmed earlier work – children gave significantly more logical responses than adults in underinformative scenarios. However, in experiment 2, when confronted with underinformative statements both adults and children were sensitive to underinformativeness as both groups preferred to choose the intermediate option (i.e. the medium strawberry) to reward Mr. Caveman. Choosing the intermediate options demonstrated that participants could tell that 'something is odd about this statement'. Essentially, they noticed the pragmatic violation. Importantly, in experiment 2, the authors found no significant difference between the groups' answer scores – indicating no evidence of a difference in pragmatic competence between groups.

This suggests that by providing three options (or more), participants are not forced to make the metalinguistic decision to reject the input or not. Instead, they can demonstrate their pragmatic competence by choosing intermediate answer options (Veenstra & Katsos, 2018). Accordingly, graded tasks are sufficient to investigate pragmatic competence in the form of *sensitivity to underinformativeness* (ib.). Importantly, Katsos and Smith (2010) note that even with this measure it is not possible to know for certain whether participants' reactions stem from (a) the derivation of scalar implicatures (which requires an initial sensitivity to underinformativeness) or (b) merely from sensitivity to underinformativeness. For example, in both case (a) (participant notices underinformativeness and derives implicature) and (b) (participant only notices underinformativeness) Mr. Caveman's initial use of *some* in the underinformative stimulus-statement is suboptimal and warrants intermediate answers (Katsos & Bishop, 2011; Veenstra & Katsos, 2018). Critically, graded SJTs (see Table 1) allow distinctions of pragmatically competent (yellow and red fields) and incompetent (green field) participants concerning their sensitivity to underinformativeness. In contrast, binary SJTs (see Table 1) only allow distinctions of pragmatically tolerant (green field) and intolerant participants (red field) concerning the tolerance of their metalinguistic judgements. Importantly, regarding pragmatic competence, Katsos and Bishop's (2011) study demonstrated that their participants were (at least) sensitive to underinformativeness (because they did not categorically accept pragmatic violations) and that there were no differences between the adult and child groups in this respect.

Similarly, Katsos and Smith (2010) tested children's pragmatic abilities by employing five-point Likert scale SJTs and yielded results similar to Katsos & Bishop (2011). Although they do not state why they used quinary scales as opposed to ternary scales, they found again that, when given the opportunity, children clearly demonstrated their sensitivity to underinformativeness by choosing intermediate options. Likewise, Jasbi et al. (2019) gave groups of adult English monolinguals either binary or graded judgement options for the same underinformative input. For example, they presented an underinformative sentence such as *There is a cat or a dog* relative to a visual stimulus depicting a cat *and* a dog. Participants were randomly assigned to one condition (either two, three, four or five decision-options SJTs). They demonstrated

that participants in the binary condition tended to categorically accept underinformative statements, while participants in graded conditions gave nuanced responses, even though in theory they could have chosen the most extreme categorical options. The authors argue that the hypothesized link between binary agree/disagree SJT paradigms on the one hand and participants' pragmatic competence on the other hand is insufficient as binary SJTs lack sensibility to pragmatic nuances such as sensitivity to underinformativeness. According to the authors, prior experimental inquiry investigating such links between response behaviour and implicature derivation rate has been severely flawed. In light of such criticism, Katsos and Smith (2010) summarize that what is different between participant groups in binary SJTs is their metalinguistic disposition relative to underinformative stimuli – not their pragmatic competence. Methodologically speaking, graded SJTs test pragmatic competence (i.e. sensitivity to underinformativeness) while binary SJTs test pragmatic tolerance (Katsos, 2021: personal communication).

2.4.2 Pragmatic Tolerance Hypothesis

In light of the different results of binary and graded SJT experiments, Davies and Katsos (2010) were among the first to establish the Pragmatic Tolerance Hypothesis. It predicts that children have pragmatic competence and what develops with age is their disposition towards underinformative items (i.e. their metalinguistic attitudes) (Davies & Katsos, 2010). Essentially, this means that the difference in answer behaviour between children and adults stems from different levels of metalinguistic awareness and attitude – not from pragmatic competence.

Veenstra et al. (2017) investigated the Pragmatic Tolerance Hypothesis empirically by testing Dutch children's (n=75, mean age: 6;3) reaction towards Dutch underinformative *some* statements in two experiments – a binary and a ternary SJT – with the same child participants in each experiment. Results indicated that 50% of the children who accepted underinformative statements in the binary SJT chose the 'intermediate' option in the ternary SJT. According to the authors, those children were pragmatically competent because they indicated sensitivity to underinformativeness in the ternary SJT and they were also pragmatically tolerant because they accepted the statement in the binary SJT. Importantly, the level of analysis of binary data has now

shifted from ‘pragmatic competence and implicature derivation ability’ to ‘pragmatic tolerance and metalinguistic awareness’. Katsos and Smith (2010), Katsos and Bishop (2011) and Jasbi et al. (2019) (discussed above) also interpreted their data in light of the Pragmatic Tolerance Hypothesis.

In sum, Katsos and colleagues’ main argument has been that binary SJTs are insufficient instruments to determine participants’ implicature derivation abilities (e.g. Veenstra & Katsos, 2018). By providing more decision options, participants are not forced to make the metalinguistic decision whether to reject the input or not, instead they can indicate that they are sensitive to underinformativeness by choosing an intermediate answer option.

In the L2 research context to date, all studies comparing scalar implicature interpretation of native speakers to L2 learners employed binary SJTs (e.g. Dupuy et al., 2019; Lin, 2016; Mazzaggio et al., 2021; Slabakova, 2010; Snape & Hosoi, 2018). According to the criticisms above, binary SJTs provide insufficient data regarding participants’ pragmatic competence, instead, they only measure pragmatic tolerance. I argue that criticisms from child L1 research also apply to L2 research.

2.5 Research Aim

My research aim is to apply the concept of pragmatic tolerance, which has been developed in L1 research, to L2 research. My hypothesis is that proficient L2 speakers are not more or less pragmatically competent than native speakers, but rather that the groups differ regarding their pragmatic tolerance. Following Katsos and Smith (2010) and Jasbi et al. (2019), I employ a quinary SJT and thereby aim to show that L2 learners are in fact as pragmatically competent as native speakers regarding their sensitivity to underinformativeness. In addition, I employ a binary SJT to investigate potential differences vis-á-vis the language groups’ pragmatic tolerance towards violations of informativeness. Essentially, the quinary SJT examines whether German EFL learners are pragmatically competent (i.e. sensitive to underinformativeness) compared to English native speakers and the binary SJT examines whether one of the language groups is more/less pragmatically tolerant than the other when forced to make a categorical ‘agree’/’disagree’ decision. Based on the interpretation of my results, I will

offer a reinterpretation of some previous L2 findings in terms of pragmatic competence and tolerance.

3 Methods

3.1 Research Questions and Research Design

To investigate and answer the two research questions (RQ 1+2) below, answers were elicited from participants through SJTs, supplemented by visual context (following criticisms of Guasti et al. (2005) of statements without such context, discussed in 2.3.1). There were two different measure types for the SJTs– a quinary measure, as in Katsos & Smith (2010), and a binary measure, as in Slabakova (2010). There are separate RQ's and hypotheses for data derived with each of these measures:

RQ1 – Measure type 1 (quinary measure):

Are German EFL learners sensitive to underinformativeness?

H_0: German EFL learners are not sensitive to underinformativeness.

H_1: German EFL learners are sensitive to underinformativeness.

RQ 2 – Measure type 2 (binary measure):

Is one of the groups of German EFL learners and English L1 speakers more tolerant of pragmatically infelicitous statements than the other?

H_0: Neither group is more tolerant of pragmatically infelicitous statements than the other.

H_{1A}: German EFL learners are more tolerant of pragmatically infelicitous statements than English L1 speakers.

H_{1B}: English L1 speakers are more tolerant of pragmatically infelicitous statements than German EFL learners.

Separate analyses were conducted to address these two RQs. For both questions, language-group (English L1 or German EFL) was the critical independent variable. Both groups were tested in English only. For the German EFL group, this ensured that bias which might result from being tested on the same matter in one's L1 and L2 during the same experiment were eliminated (see criticisms of Dupuy et al., 2019; discussed in 2.3.1). In addition, there were two different versions of the experiment: (1) one version with the quinary measure only and (2) one version with the binary measure

only. Each individual participant from each language group was randomly allocated to one version. Testing each participant in either the quinary measure or the binary measure eliminated learning effects or bias which might have resulted from being tested in both measure types during the same experiment.

While underinformative *some* statement-types were the targets, control sentences (optimally true *all*, optimally false *all*, felicitous *some*) were included as well, following Slabakova (2010). For these, all participants were expected to show the same (ceiling) performance. These control-data did not contribute to answering the hypotheses and were analysed separately. Control items ensured that results obtained from target items were due to the linguistic variable under examination (viz. informativity) and not due to other issues in the task structure (cf. Schmitt & Miller, 2010).

3.2 Participants

This study targeted native speakers of English and German native speakers learning English at school. All participants were high-school students and at least 16 years old. According to criteria in the ethics approval, participants were considered 'competent youths', hence, they gave their own consent and did not need their parents' or legal guardians' assent to participate. In total, 66 participants took part in the study. 35 participants completed the quinary task (German EFL n=19; English L1 n=16) and 31 participants completed the binary task (German EFL n=17; English L1 n=14). As suggested by Dörnyei (2007), a priori power analyses using G*Power Version 3.1 (cf. Faul et al., 2009) were conducted for both the binary and quinary measures. Results recommended a slightly larger sample size (minimum n=19 per language group) for the binary SJT and a larger sample size (minimum n=35 per language group) for the quinary SJT to accurately detect significant effects in within-group and between-group tests of mean difference, respectively (see 5.4). Detailed power analyses are provided in Appendices 1a and 1b. Unfortunately, it was not possible to recruit this sample with the resources and time frame available; implications from the fact that the study was underpowered are addressed in the 5.4.

3.2.1 Sampling

The sample of German EFL learners was drawn from a selective grammar school in Germany. They were studying English in an instructed learning and non-immersion educational context. The learners were in years ten to twelve (i.e. aged between 16-18), equivalent to a British 'Sixth Form'. Although the sample was not representative of the German society as a whole, it was a convenient sample for the study's purpose, and it was used for convenience of recruitment because the researcher had previously visited and had personal contacts with the school. Recruitment from a school instead of from the internet also facilitated the control for homogeneity of both participant characteristics and proficiency.

The German EFL learners had studied English for at least eight years and had sufficient knowledge of English to partake in the study. According to the official language policy of the Ministry of Education of the Federal State of Baden-Württemberg, their English proficiency conformed to B2-C1 in the CEFRL (cf. Kultusministerium Baden-Württemberg, 2016). Each German participant confirmed prior to the experiment that they are monolinguals.

The sample of English native speakers was drawn from a non-selective British independent senior school. Although this was not representative of the British society as a whole, it was a convenient sample for the study's purpose, as this group was age-matched to the German students. Additionally, the sample was used for convenience of recruitment because the researcher used to work at the school. As for the German group, drawing the samples from a school instead of from the internet facilitated the control for homogeneity of participant characteristics.

Since the researcher was well acquainted with both schools it could be confirmed from personal experience that the demographics at both schools were similar. This was important to ensure that socio-cultural influences on the interpretation processes (i.e. on the participants' responses) during the experiment could be minimized.

3.3 Instruments

This study used the online experiment platform *Gorilla* (www.gorilla.sc; cf. Anwyl-Irvine et al., 2020). It provided appropriate technical response-recording requirements for this study's purpose (Anwyl-Irvine et al., 2020). Participants accessed

the experiment via a link. They were presented with an online participant information sheet and gave their consent. Only after they had given their consent, they could access the experiment.

3.4 Stimuli

Both measure types provided visual contexts. Each experimental item consisted of a visual two-pictures kitchen scenario and a statement describing the scenario and containing either the quantifier *some* or *all*. The two pictures showed distribution activities in everyday kitchen settings. For example, the first picture showed a set of nine carrots on a table next to a plate. In the second picture, the plate had now some carrots on it (see Figures 4 and 5 below). The accompanying statement described the final state in either an optimally true, optimally false, pragmatically felicitous, or underinformative manner. This was expressed via a sentence such as *The chef put all of the carrots on the plate*. Beneath the experimental item, participants were provided with buttons where they chose either between 'disagree' / 'somewhat disagree' / 'neither' / 'somewhat agree' / 'agree' (quinary mease) or 'agree' / 'disagree' (binary measure) options. A blue progress bar was provided next to the experimental items (see Figures 4 and 5 below). Each set of kitchen equipment or groceries consisted of nine items (i.e. there were always nine carrots/lemons/cups in total and never eight, seven, etc.). Although other studies have used smaller sets (e.g. Dupuy et al., 2019), Snape and Hosoi (2018) noted that small sets make scalar scales such as <some, all> more inconclusive and impact interpretational behaviours. Each kitchen scenario was combined once with each of the four statement types (see Fig.6).

Fig. 4 Screenshot of an experimental trial in the quinary measure (here: optimally false all sentence)

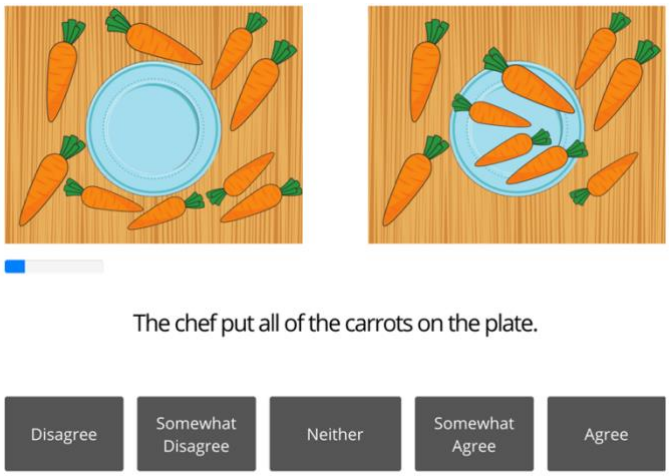


Fig. 5 Screenshot of an experimental trial in the binary measure (here: optimally false all sentence)

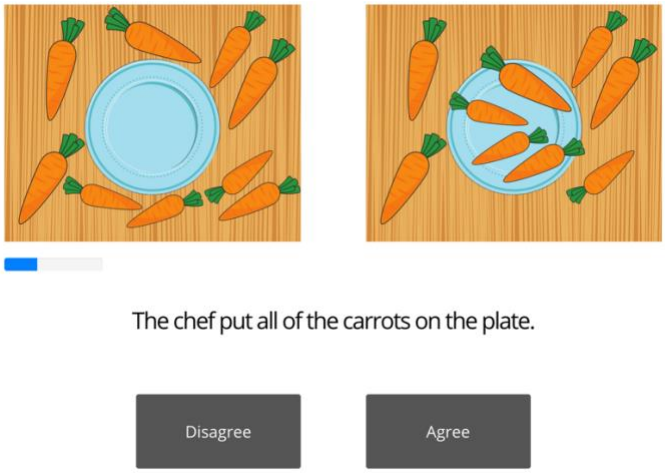
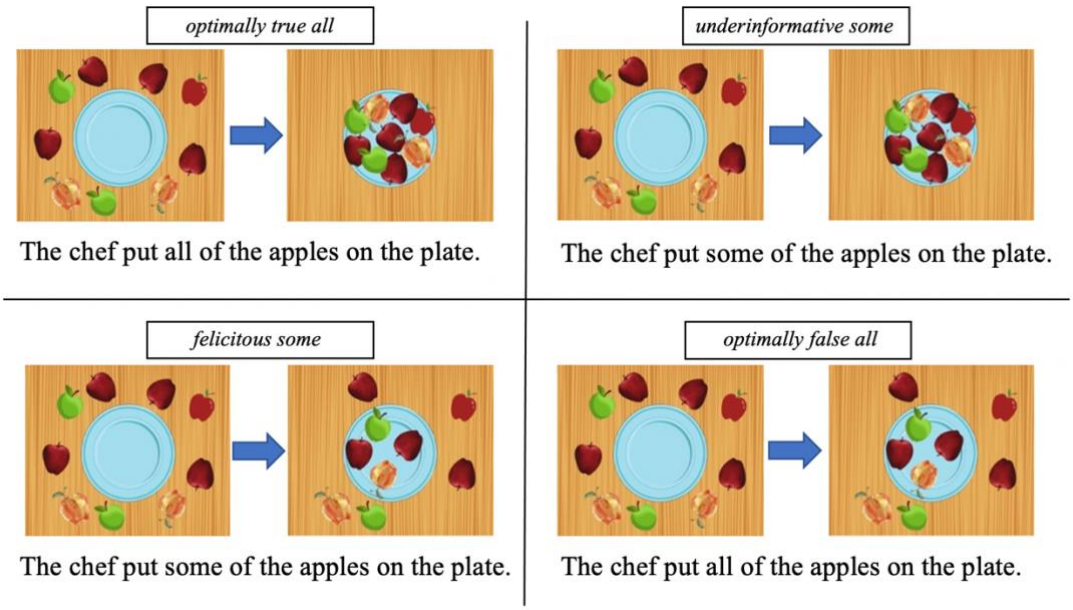


Fig. 6 Example of one kitchen scenario (here: apples) combined with all four statement types.



In total, twelve different kitchen scenarios were created (e.g. carrots, apples, yoghurts, mushrooms, etc.), therefore, the overall number of experimental items added up to 48 (twelve types of kitchen scenarios *x* four statement-types; Figure 6 is an example of one of the kitchen scenarios combined with all four statement types). Each sentence was proof-read by English L1 speakers. Additionally, the sentences were evaluated by a Modern Foreign Language teacher in the German grammar school who confirmed that the stimuli were at an appropriate language level for the German EFL learners. The underlying ditransitive sentence structures was the same across both target and control sentences: they began with an NP in subject position, followed by a VP including a direct object NP and a PP. All items had approximately the same length. Controlling sentence properties eliminated syntactical factors as a source of influence on participant-behaviour (cf. Schmitt & Miller, 2010). More example experimental items are provided in Appendix 2.

3.5 Procedure

RQ1 was addressed via asking participants to give responses on a five-point Likert-scale (‘disagree’ / ‘somewhat disagree’ / ‘neither’ / ‘somewhat agree’ / ‘agree’) to underinformative *some* sentences. RQ2 was addressed by asking them to give binary responses (‘disagree’ / ‘agree’) to underinformative *some* sentences.

Participants completed the experiment at home on their personal electronic device (e.g. computer, laptop, tablet). They began the experiment by viewing the digital participant information sheet which provided detailed information about the study (see Appendix 3). At the end of the participant information sheet, participants were thanked for their collaboration and asked to complete an online consent form on the next screen (see Appendix 4). Both documents were displayed in English for the English L1 participants and in German for the German EFL participants. Participants were not explicitly informed that they would be tested on scalar terms as this might have influenced their behaviour. As soon as they had given their consent, they were told that they would view a series of kitchen-scenario-pictures and statements and that with each picture-statement combination they would be given answer choices which they could select by clicking on them. *Gorilla* always displayed the answer options from left to right in the order 'disagree' / ('somewhat disagree' / 'neither' / 'somewhat agree') / 'agree' (see Figures 4 and 5 above).

Gorilla randomly assigned 50% of participants of each language group to either the quinary measure pathway or the binary measure pathway. The randomisation mode was balanced, to ensure that the same number of participants (per language group) was allocated to each pathway. This was important for statistical analyses.

The participants were introduced to a scenario where a friend who they had met last summer (whose first language is not English) visits and they watch a cooking show together. The chef in the show moves kitchen equipment and groceries around and each time the chef has moved something, the friend comments on it, making statements such as *The chef moved all of the eggs in the bowl*. Since the friend is not proficient English speaker, s/he sometimes makes mistakes. The participants' task was to evaluate the friend's statements. They were asked to answer as fast as possible. The instructions are available in Appendix 5. Three items were presented for training purposes. Then, once the participant pressed 'Start', the experiment began.

Each pathway included 48 trials. A trial consisted of one experimental item (defined above, see 3.4). The experimental items were the same for each pathway and the order of trials was randomized for each participant. Between each trial, the screen was empty for 1000ms. In both pathways, the answer options always appeared on the screen at the same time as the experimental item (see Figures 4 and 5 above). Once

finished, the participants were thanked for their participation. They received a short debriefing about the experiment's purpose. The experiment took about five minutes. Critically, participants encountered each experimental item (i.e. statement-scenario combination) only once throughout the entire experiment to avoid learning-effects, and each pathway (binary and quinary) contained the same experimental items to allow for post-hoc comparisons between conditions (Katsos & Bishop, 2011).

3.6 Ethical Considerations

Ethics approval was obtained from the Departmental Research Ethics Committees (DRECs) of the Department of Education at the University of Oxford (for ethics approval see Appendix 6; Reference-No.: ED-CIA-21-164).

Following the Central University Research Ethics Committee's regulations on Research Involving Competent Youths, the participants were considered 'competent youths' as "'competent youths' are aged 16 to 17" (University of Oxford CUREC, Best Practice Guidance 04_Version 2.3, 2020) Therefore, it was assumed that all the students "have sufficient understanding of the project and its implications for them that they can make up their own minds about taking part, and have that opinion honoured" (ib.). Corresponding to section 7, point viii. in the Best Practice Guidance 04_Version 2.3, "examples of research where the class of 'competent youths' would generally apply [include] studies of language use, perception or production" (ib.). This study has been approved as such a project involving competent youths. Therefore, parental consent was not required.

Both schools received an invitation letter from the researcher. They also received copies of the online participant information sheet, the online consent form and the CUREC approval. On the online consent form, the participants had to confirm explicitly that they are at least 16 years old. The participants were informed that participation is voluntarily and that they could withdraw from the experiment at any time without providing reasons. In that case, their data would be deleted immediately. Moreover, the participants were told that in case of any questions or concerns they could contact the researcher or the supervisor.

3.7 Pilot

The study was piloted among some of the researchers' friends. After the pilot, two methodological changes were implemented. First, time limits during trials were removed as participants tended to answer randomly under time pressure. Additionally, it was found that relevant literature did not sufficiently justify time limitations in this type of experiment. Second, progress bars were introduced to show participants how much longer they would have to focus.

4 Results

This chapter reports quantitative results of the present study; data analysis is ordered by research question. First, an analysis of the quinary data is conducted to answer RQ1 ('Are German learners of English sensitive to underinformativeness?'), followed by an analysis of the binary data for RQ2 ('Is one of the groups of German EFL learners and English L1 speakers more tolerant of pragmatically infelicitous statements than the other?'). All analyses were run using IBM SPSS version 27 statistical software.[3]

4.1 Sensitivity to Underinformativeness – Pragmatic Competence (RQ1)

4.1.1 Scoring quinary data

The answer options in the quinary measure were 'disagree' / 'somewhat disagree' / 'neither' / 'somewhat agree' / 'agree'. To conduct statistical analyses, participants' answers for the item types optimally true *all*, felicitous *some* and underinformative *some* were scored numerically from 1 to 5 as illustrated below:

Disagree	-	*Somewhat Disagree*	-	*Neither*	-	*Somewhat Agree*	-	*Agree*
1		*2*		*3*		*4*		*5*

Following Pipijn and Schaeken (2012), participants' answers for the optimally false *all* items were scored in reversed order, i.e.

Disagree	-	*Somewhat Disagree*	-	*Neither*	-	*Somewhat Agree*	-	*Agree*
5		*4*		*3*		*2*		*1*

As a result of this scoring, participants' mean answer scores for the three control item types (optimally true *all*, optimally false *all*, felicitous *some*) were expected to approximate '5' because participants normally agree with optimally true *all* and felicitous *some* items and disagree with optimally false *all* items. Regarding the target underinformative *some* items, the closer the score approximates '5', the more *logical* the mean answer behaviour is and the closer the score approximates '1', the more *pragmatic* the mean answer behaviour is.

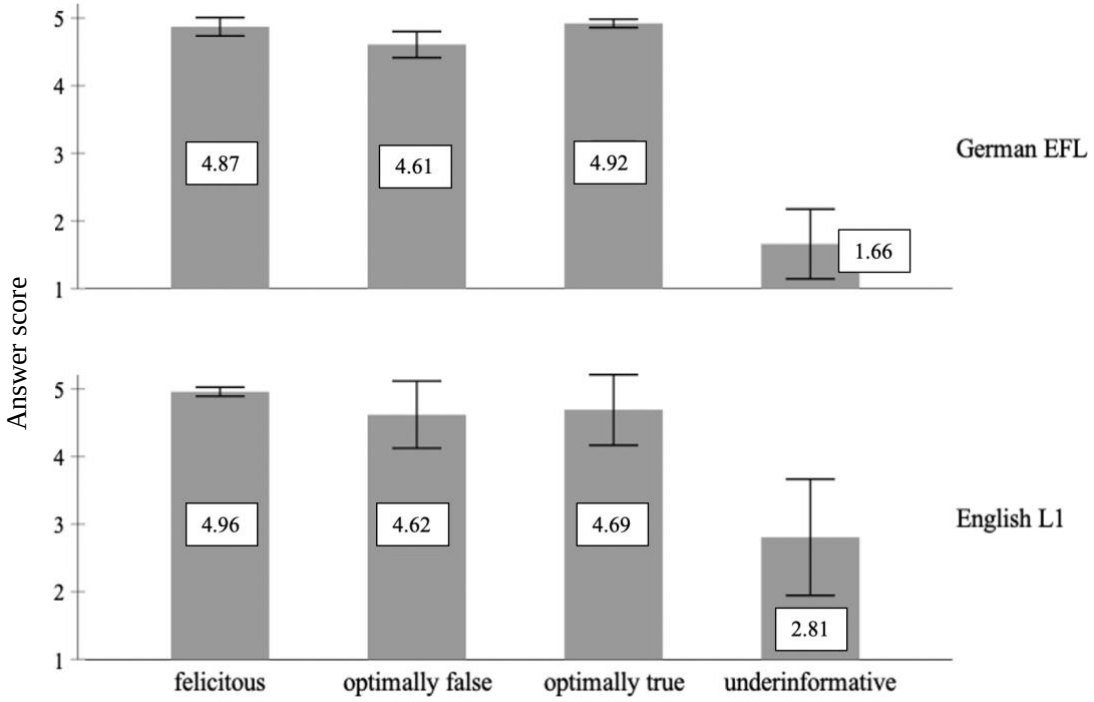
[3] IBM Corp. 2020. IBM SPSS Statistics for Macintosh, Version 27.0. Armonk, NY: IBM Corp.

4.1.2 Analysis

RQ1 inquires whether German EFL learners are sensitive to underinformativeness. As discussed in 2.4, only the quinary measure gives participants the opportunity to unambiguously demonstrate their sensitivity to underinformativeness. If participants in the quinary measure gave intermediate answers (i.e. 'somewhat disagree', 'neither' or 'somewhat agree') to underinformative *some* items or disagreed (i.e. 'disagree') with underinformative *some* items, they were considered 'sensitive to underinformativeness'. Statistically, we can evaluate this hypothesis by seeing if they give lower scores to these underinformative items than they do with the control items, where they should have scores near '5' (i.e. ceiling).

Results are presented in Figure 7. On visual inspection, both the English L1 group and the German EFL group performed at ceiling (i.e. approximating '5') in all control types, meeting the expectations for participants who are rational language users without visual impairments (cf. Dupuy et al., 2019).

Fig. 7 Mean answer score of all participants in each language group per item type in the quinary measure. Means and 95% confidence intervals are displayed.



In contrast, the mean answer scores for underinformative items in each language group appear lower than the mean scores for the control items. To investigate this observation statistically, a within-groups test of mean difference was conducted for each of the language groups: a non-parametric test was used given the small sample size (Hatch & Lazaraton, 1991). Additionally, skewness values (German EFL/English L1: optimally true *all* = -1.42/-3.94; optimally true *false* = -1.09/-3.33; felicitous *some* = -3.16/-3.71; underinformative *some* = 2.0/.161) indicated non-normality of distribution and Shapiro-Wilk tests of normality were all < *.05*, suggesting that the data were not normally distributed, supporting the use of a non-parametric test (Field, 2018).[4] The German EFL group's optimally true *false* data and the English L1 group's underinformative *some* data were examined visually since their skewness values were somewhat weak at -1.09 and .161, respectively (see Appendix 7, Figures A17 and A18). The visual examination also supported the use of a non-parametric test. Therefore, a Friedman's ANOVA was run for each language group. Results indicated that in both language groups there was evidence for significant differences in the participants' mean response scores depending on statement type (English L1: Friedman's ANOVA $\chi^2(3) > 22.525$, $p < .05$; German EFL: Friedman's ANOVA $\chi^2(3) > 42.865$, $p < .05$), thus, post-hoc analyses with Wilcoxon signed-rank tests were conducted. Since the sample sizes were small (below n=50), exact significance values were calculated in SPSS instead of asymptotic values (cf. Field, 2018).[5]

Post-hoc Wilcoxon signed-rank tests revealed that in the German EFL group there were no significant differences in answer scores between the optimally *true* and felicitous *some* statement types. However, there were significant differences in answer scores between all the remaining pairs, namely, between 'optimally false *all* – felicitous *some*' ($Z = -2.005$, $p < .05$, $r = .33$), 'underinformative *some* – felicitous *some*' ($Z = -3.833$, $p < .05$, $r = .62$), 'optimally true *all* – optimally false *all*' ($Z = -2.870$, $p < .05$, $r = .47$), 'underinformative *some* – optimally false *all*' ($Z = -3.712$, $p < .05$,

[4] Note that given the small sample sizes (n=19 and n=16), the Shapiro-Wilk tests may well be underpowered (Field, 2018).

[5] To avoid rejections of the null hypothesis based on false assumptions, Bonferroni correction was applied to the Wilcoxon signed-rank tests' p-values (Field, 2018). However, the calculations resulted in no noteworthy changes to the significance values, therefore, the corrections are not reported. In addition, effect sizes (r) were calculated manually by dividing the absolute standardized test statistic Z by the square root of the number of total observations (Rosenthal, 1991).

r = .60) and 'underinformative *some* – optimally true *all*' (Z = -3.830, p < .05, r = .62). While the significant differences in pairs including underinformative *some* statement types were expected, the results from the 'optimally false *all* – felicitous *some'* and the 'optimally true *all* – optimally false *all*' pairs were unexpected (discussed in 5.1.1). However, effect sizes in the pairs containing *underinformative* items are stronger (r ≥ .6) compared to the other pairs (r < .5). According to Muijs (2011), r ≥ .6 is considered a moderate to strong effect, while r ≤ .5 is considered a modest effect. Essentially, the significant differences in the pairs containing underinformative *some* items are larger compared to the significant differences in the pairs without underinformative *some* items.

In comparison, in the English L1 group there were no significant differences in answer scores between the 'optimally false *all* – felicitous *some*', 'optimally true *all* – felicitous *some*' and 'optimally true *all* – optimally false *all*' statement type pairs. However, there were significant differences in mean answer scores between all the remaining pairs, all of which included underinformative *some* statement types, that is, 'underinformative *some* – felicitous *some*' (Z = -3.190, p < .05, r = .56), 'underinformative *some* – optimally false *all*' (Z = -2.703, p < .05, r = .48), 'underinformative *some* – optimally true *all*' (Z = -3.830, p < .05, r = .46).

In sum, regarding RQ1, it can be said that there was evidence that both groups showed sensitivity to underinformativeness as they rated underinformative *some* items significantly lower than the control items.

In addition to answering RQ1, it seemed worthwhile to conduct between-groups tests of mean answer score differences to compare the German EFL group's answer behaviour to the English L1 group's behaviour. Therefore, four Mann-Whitney U-tests (between-group comparison for non-parametric data), with 'language group' as grouping variable and 'statement type' as dependent variable, were conducted following Katsos and Bishop (2011). For all four statement types, the tests revealed no significant differences in answer behaviour between language groups (felicitous *some*: U = 129.5, p > .05; optimally false *all*: U = 110.0, p > .05; optimally true *all*: U = 144.0, p > .05; underinformative *some*: U = 96.0, p > .05). This indicates that there was no evidence that the two language groups differed in their responses to the four statement types, including for the items testing underinformativeness.

Note that Figure 7 only illustrates mean answer scores at group level without providing detailed information as to *how often* each decision option was chosen per statement type and language group. To gain more clarity regarding participants' detailed answer behaviours, mean proportions of answers to all four statement types (see Table 2) were inspected to see if participants indeed were giving intermediate answers ('somewhat agree' / 'neither' / 'somewhat disagree') to underinformative items rather than having the means just represent a mix of categorical 'disagree' and 'agree' responses.

Table 2. Mean proportion of answers to all four statement types per language group in the quinary measure.

statement type	**German EFL**	**English L1**
felicitous		
Disagree	0	0
Somewhat Disagree	0	0
Neither	0.01	0
Somewhat Agree	0.04	0.03
Agree	0.95	0.97
optimally false		
Disagree	0.86	0.91
Somewhat Disagree	0.11	0.07
Neither	0.01	0
Somewhat Agree	0.01	0.01
Agree	0.01	0.01
optimally true		
Disagree	0	0.01
Somewhat Disagree	0	0
Neither	0	0
Somewhat Agree	0.02	0.02
Agree	0.98	0.97
underinformative		
Disagree	0.44	0.14
Somewhat Disagree	0.12	0.11
Neither	0.03	0
Somewhat Agree	0.24	0.25
Agree	0.17	0.50

As shown in Table 2, it was indeed the case that participants gave intermediate answers to underinformative items, whereas they gave mostly categorical responses to

the control items (as expected). Furthermore, the median answer scores for underinformative *some* items in each language group were compared to the mean scores (German EFL: M = 1.66, MD = 1.33; English L1: M = 2.81, MD = 2.62). Since the two scores were similar in each language group, this comparison indicated that in fact the mean answer scores at group level were not inherently split into two groups (i.e. participants with high mean answer scores and participants with low mean answer scores). Additionally, the answer behaviour data (for underinformative items) was analyzed at the level of individual participants. Thereby, each participant was put in one of three categories: *consistent sensitivity to underinformativeness* (i.e. they gave only 'disagree', 'somewhat disagree', 'neither' or 'somewhat agree' answers), *inconsistent sensitivity to underinformativeness* (i.e. they used the entire range of answer options, including 'agree') or *no sensitivity to underinformativeness* (i.e. they gave only 'agree' answers) (see Table 3).

Table 3. Individual participants' consistency regarding their answer behaviour towards underinformative items in the quinary measure. Number (and percentage) of participants per language group are presented.

	Consistent sensitivity to underinformativeness	**Inconsistent sensitivity to underinformativeness**	**No sensitivity to underinformativeness**
German EFL (n=19)	12 (63.2%)	7 (36.8%)	0 (0.0%)
English L1 (n=16)	8 (50.0%)	5 (31.2%)	3 (18.8%)

As demonstrated in Table 3, most participants in each language group were considered consistently sensitive to underinformativeness. Combined with the analysis of the relationship between the median and mean scores, this indicates that the language groups were not inherently split into pragmatically competent and pragmatically oblivious individuals, and that participants were indeed giving intermediate answers to underinformative items.

Although the language groups' mean answer scores for underinformative items were similar (see Figure 7), it was not clear whether the detailed answer patterns differed between language groups. As demonstrated in Table 2, there may be some differences between the groups: while the German EFL group gave more categorical

'disagree' answers (44%) than the English L1 group (14%), the native speakers seemed to categorically agree with underinformative *some* items more often (50%) than the non-native speakers (17%). However, although the 'agree' and 'disagree' answers seem to be distributed differently in each language group, one should be careful about overinterpreting those answer patterns with such small numbers where a statistical evaluation is not possible, and the distribution could be chance (cf. Field, 2018). Similarly, regarding Table 3, although the numbers of participants in each category (*consistent/inconsistent/no* sensitivity to underinformativeness) could in theory be compared using a Chi-square test (2x3 contingency table), the numbers per cell are too low to run a valid test (cf. Field, 2018). In sum, this suggests that generally the numbers are too small to be able to look at differences in answer patterns between language groups in the quinary measure.

4.2 Pragmatic Tolerance (RQ2)

4.2.1 Scoring binary data

To conduct statistical analyses regarding the binary measure, participants' answers relative to the item types felicitous *some*, optimally true *all* and underinformative *some* were scored numerically as demonstrated below:

Disagree	–	*Agree*
0		1

Like the quinary scoring, participants' answers relative to the optimally false *all* items were scored in reversed order:

Disagree	–	*Agree*
1		0

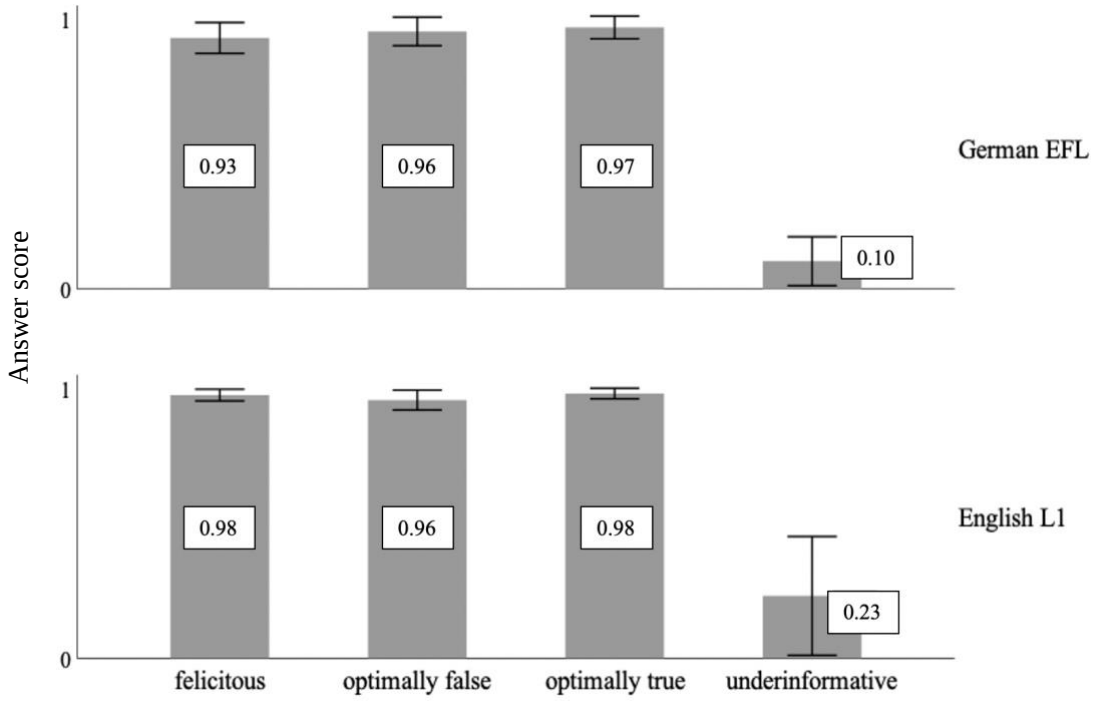
As a result of this scoring, the participants' mean answer behaviour relative to the three control item types (felicitous *some*, optimally true *all*, optimally false *all*) was expected to approximate '1' because this score constitutes the usually 'correct' reaction.

4.2.2 Analysis

RQ2 asks whether German EFL learners and English L1 speakers differ in their pragmatic tolerance: the closer the score for the critical underinformative *some* items approximated '1', the more pragmatically tolerant the participants' behaviour was

considered to be, and the closer the score approximated '0', the less pragmatically tolerant their behaviour was considered to be. In this regard, as demonstrated in Figure 8, both language groups performed at ceiling in the control conditions: they agreed to felicitous *some* and optimally true *all* statements and disagreed to optimally false *all* statements. Turning to the critical items for pragmatic tolerance: both language groups showed low pragmatic tolerance levels (German EFL M = .10, MD = .0; English L1 M = .23, MD = .04) because they disagreed with underinformative *some* statements most of the time (see Figure 8).

Fig. 8 Mean answer score of all participants in each language group per item type in the binary measure. Means and 95% confidence intervals are displayed.



To answer RQ2, a between-groups tests of mean difference was conducted to investigate whether there was a significant difference between the two language groups' mean answer behaviour for underinformative *some* items in the binary measure: Following Katsos and Bishop (2011), and given the non-normality of the data (based on results of Shapiro Wilk tests for normality (German EFL: $W(17) = .649$, $p < .05$; English L1: $W(14) = .643$, $p < .05$), on the data's skewness values for underinformative statement types (German EFL = 2.369; English L1 = 1.533) and on a visual examination of histograms (see Appendix 8, Figures A19 and A20)), a non-

parametric Mann-Whitney U-test with 'language group' as grouping variable was conducted. This revealed no significant difference in mean answer behaviour towards underinformative *some* items between the two language groups ($U = 105.5$, $p > .05$). Thus, there was no evidence to suggest that the German EFL group was more or less pragmatically tolerant than the English L1 group.

As a further analysis, similar to Slabakova (2010), the underinformative *some* data were re-analyzed by splitting each language group into three subgroups: participants who disagreed over 80% of the time, participants who agreed over 80% of the time and participants who made unclear choices, that is, they agreed or disagreed fewer than 80% of the time and, showing inconsistent behaviour – this is shown in Table 4.

Table 4. Three types of answer behaviour relative to underinformative items in the binary measure. Number (and percentages) of pragmatically intolerant/tolerant/unclear participants per language group are presented.

	Number of participants that provided		
	pragmatically intolerant (i.e. 'disagree') answers over 80% of the time	**pragmatically tolerant (i.e. 'agree') answers over 80% of the time**	**unclear answer patterns**
German EFL (n=17)	15 (88.2%)	0 (0.0%)	2 (11.8%)
English L1 (n=14)	10 (71.4%)	2 (14.3%)	2 (14.3%)

To see whether there was an association between language group and disagreement with underinformative *some* items over 80% of time (i.e. low pragmatic tolerance), a Chi-square test of association was performed. To avoid empty cells, as recommended by Field (2018), the 'number of participants who agreed over 80% of the time' and the 'unclear choices participants' were grouped together to act as categorical counterpart to the 'participants wo disagreed over 80%' of the time (see Table 5).

Table 5. 2x2 contingency table of 'language group' and 'disagreed over 80% of the time'.

	Number of participants who disagreed over 80% of the time		
	No	**Yes**	**Total**
German EFL	2	15	17
English L1	4	10	14
Total	6	25	31

Given the small sample sizes, the Fisher's Exact Test's p-value is reported as suggested by Field (2018) for 2x2 contingency tables with expected frequencies below 5 per cell. The test's null hypothesis is that the two categorical variables are independent (i.e. *language group* and *disagreed over 80% of the time*). Based on the results of the 2x2 Chi-square analysis (χ^2 (1) = .3697, p > .05), the null hypothesis was retained. The analysis thus revealed no evidence that there was a difference between the language groups in terms of the proportion of participants categorized as having/not having low levels of pragmatic tolerance.

5 Discussion

The current work explored L2 learners' pragmatic competence and tolerance by looking at German EFL learners' responses to underinformative *some* statements using binary and quinary measurements, and by comparing their performance to the performance of English L1 speakers on the same tasks. In the quinary measure both language groups had statistically significant lower mean scores on underinformative *some* items than they did on the control sentences, reflecting the fact that they gave more intermediate responses for underinformative items than for control items. Moreover, looking at individual participants, it was found that most German EFL and English L1 participants demonstrated pragmatic competence. There was no evidence for between-group differences regarding answer-behaviour towards underinformative items. In the binary measure both groups had low mean scores for the underinformative *some* sentences, reflecting the fact that they tended to reject underinformative items, indicating that levels of pragmatic tolerance were low in both language groups. There was no evidence for between-group differences, either when comparing mean scores or comparing numbers of participants categorized as having/not having low levels of pragmatic tolerance.

In the following discussion, findings concerning the quinary measure are discussed first, followed by a discussion of the findings from the binary measure. Lastly, in the context of both discussions regarding the quinary and the binary data, the unexpected findings that (a) some participants in the quinary measure appeared to be pragmatically oblivious and (b) some participants in both measure types showed inconsistent behaviours will be addressed.

5.1 L2 Learners' Pragmatic Competence (quinary measure)

The data reported in this thesis suggests that the German EFL learners can be considered pragmatically competent – at least to the extent of sensitivity to underinformativeness (cf. Veenstra & Katsos, 2018). Therefore, the current thesis's findings with L2 learners are in line with Katsos and Bishop's (2011) results with L1 children (using a similar ternary judgement task) in that participants demonstrated sensitivity to underinformativeness because they did not categorically accept items which represented pragmatic violations. Simply put, participants consistently noticed

that the underinformative *some* items were 'odd'. This finding is the core finding of the current thesis as it unambiguously provides evidence for L2 learners' pragmatic competence based on a quinary SJT – unlike some previous L2 studies using binary SJTs. Furthermore, conforming to Katsos and Bishop's (2011) results with children and adults, there was no significant difference in mean answer behaviour towards underinformative *some* items between the English L1 group and the German EFL group, suggesting no evidence that English L1 speakers and German EFL learners differed in their sensitivity to underinformativeness.

These findings differ from the conclusions drawn by researchers in previous L2 studies regarding the pragmatic competence of L2 learners compared to native speakers. Those studies, critically, all used binary judgement tasks. For example, as discussed in 2.3.1, Mazzaggio et al.'s (2021) data indicated that L2 learners accepted underinformative *some* statements more often than native speakers (i.e. logical response; *some and possibly all*), which they interpret as showing that L2 speakers are less likely to derive scalar implicatures compared with native speakers. Slabakova (2010) makes a similar reverse argument because her data indicates that L2 learners are *more* likely to derive the implicature (*some but not all*) than L1 speakers. In comparison to such conclusions, the quinary measure data from the current thesis justifies more nuanced theoretical claims: L2 learners are sensitive to underinformativeness and there is no evidence that they are less sensitive to underinformativeness than native speakers.

With regards to the between-group comparisons, it is important not to overinterpret the null-result (i.e. no evidence for a between-group difference regarding sensitivity to underinformativeness) with such a small study. The study could be underpowered to detect a difference (see 5.4). Nevertheless, a tentative reason as to why the groups did not differ might pertain to the participants' cognitive abilities: the L2 learners were reasonably proficient (B2, CEFRL), the input statements were simple, and the visual contexts were straightforward. The pilot study with friends and feedback from the L2 learners' English teacher before and after running the experiment confirmed that the L2 participants should not (and did not) encounter any problems understanding the input. Therefore, according to Kecskes (2019), there is no reason from a strictly cognitive abilities perspective to assume that normally developed L2 learners without

learning impairments differ fundamentally from native speakers in terms of general pragmatic abilities, including the ability to detect basic pragmatic violations. This could explain why there was *no* difference found between language groups in terms of their sensitivity to underinformativeness in the current study. However, it must be acknowledged that the same participant characteristics would be true for participants in studies that *find* differences between groups in quinary tasks, although such findings have not been reported yet (to the best of the researcher's knowledge).

Recall from 2.4 that Veenstra and Katsos (2018) have argued that even graded judgement scales cannot provide direct evidence that participants derive implicatures and that this requires other experimental paradigms, for example eye-tracking techniques or EEG techniques that capture more exact time-courses of language processing and, therefore, provide insights into different processing stages of implicature derivation (e.g. Huang & Snedeker, 2009; Noveck & Posada, 2003; Yoon et al., 2015). For example, the N400s in Noveck and Posada's (2003) ERP study (discussed in 2.2.3) showed that there was no immediate reaction to underinformative sentences in the participants' brains, which allowed the researchers to make assumptions about the time-course and triggers of implicature processing. Similar research in the realm of L2 studies might be promising concerning L2 cognitive implicature derivation processes.

5.1.1 German EFL within-group differences

One unexpected finding in the quinary measure data was that the German EFL learners' answer behaviour differed significantly in the 'optimally false *all* – felicitous *some*' and the 'optimally false *all* – optimally true *all*' pairs.[6] Although differences were small, participants in the German EFL group chose the 'somewhat disagree' option in the optimally false *all* scenarios (11% of all answers) significantly more often compared to its (reversed) counterpart, the 'somewhat agree' option in both the

[6] Recall examples:

Optimally false *all*:	*The chef put all of the bananas on the table* in a context where only some of the bananas are on the table.
Felicitous *some*:	*The chef put some of the bananas on the table* in a context where some of the bananas are on the table.
Optimally true *all*:	*The chef put all of the bananas on the table* in a context where all of the bananas are on the table.

felicitous *some* (4% of all answers) and the optimally true *all* (1% of all answers) scenarios. Interestingly, Jasbi et al.'s (2019) quinary SJT data showed similar patterns. In their *unambiguously false* condition (equivalent to optimally false *all*), the input statement was for example *There is a cat and a dog on the card* when, in fact, there was only a cat on the card. Although participants were expected to categorically disagree with the utterance (i.e. indicate that the statement is 'wrong' in Jasbi et al.'s (2019) terms), they did so only 46% of the time. 32% of the time the participants answered that the utterance was 'kinda wrong' (i.e. equivalent to 'somewhat disagree'), that is, they avoided categorical rejections – although categorical rejections are warranted in *unambiguously false* scenarios. The authors argue that their participants might have reasoned that some statements in the experiment are 'partially true'. That means that participants could have believed that seeing a cat is partially true in a context where they should see a cat *and* a dog. A similar reasoning process could well explain the phenomenon in the current study. For example, in an optimally false *all* scenario where only some of the bananas are on the plate, but the chef utters *All of the bananas are on the plate*, the utterance could be considered partially true which would eventually provoke and justify 'somewhat disagree' answers as opposed to categorical 'disagree' answers. Additionally, as discussed above, the data suggested that German EFL participants generally noticed violations of informativeness in the underinformative *some* items. However, given that each participant evaluated 48 items in total (in random order, 12x items per statement type), it is reasonable to assume that while focusing on the task and looking for 'oddness', some participants extended their reasoning process that 'something is not quite right' (formed based on the underinformative *some* items) to the optimally false *all* items, leading them to consider optimally false *all* items to be partially true. Another way of thinking about this is that the unexpected answer behaviour could be attributed to an experimental artifact because participants were learning in the experiment that they are expected to give nuanced answers. Combined with the general oddness of underinformative *some* items, participants could have been inclined to press 'somewhat disagree' in optimally false *all* scenarios. In fact, a post-hoc comparison of German EFL learners' mean response scores to optimally false *all* items revealed that they gave higher mean scores ($M = 1.51$) in the second half of the experiment compared to the first half of the

experiment (M = 1.32). Although a paired samples t-test showed no significant difference in mean answer scores between the first half of the experiment and the second half ($p > .05$), the direction of these means is tentatively consistent with the suggestion that initially participants might have tended to give more categorical answers (i.e. 'disagree') and then began to lean more towards less categorical answers. Note that a similar interpretation in terms of partial truthfulness is not possible for felicitous *some* and optimally true *all* items (because those items cannot be considered partially true), explaining why the differences were only found in the 'felicitous *some* – optimally false *all*' and 'optimally true *all* – optimally false *all*' pairs, but not in the 'optimally true *all* – felicitous *some*' pair.

Although the English L1 group did not show this same pattern of significant within-statement-type-difference, no evidence of between-group differences for the control statement types was found and thus any interpretation of the English L1 group's answer pattern in comparison to the German EFL group's pattern discussed above should be treated as tentative. If this difference between groups (regarding the control items) did turn out to be significant in a larger sample, one possible explanation could be that the English L1 group was not tested in their L2. As suggested by Dupuy et al. (2019), being tested in L2 might temporarily increase one's metalinguistic awareness and, in the case at hand, lead to the answer pattern in the German EFL group reported above. In contrast, the English L1 participants were tested in their L1 which, following Dupuy et al.'s (2019) logic, initiated no temporary increase in metalinguistic awareness. Therefore, the English L1 group would not tend to consider optimally false *all* items to be partially true, and thus, they would not press 'somewhat disagree' as often as the L2 group in these situations.

5.2 L2 Learners' Pragmatic Tolerance (binary measure)

RQ2 investigated the language groups' pragmatic tolerance ('Is one of the groups of German EFL learners and English L1 speakers more tolerant of pragmatically infelicitous statements than the other?'). Recall, if German EFL learners were more tolerant (H1A; see 3.1) we would expect them to show a higher mean answer score compared to the English L1 group, reflecting a larger proportion of 'agree' answers to underinformative items. If the English L1 participants were more tolerant (H1B; see

3.1) we would expect them to show a higher mean answer score compared to the German EFL group.

The descriptive data suggested generally low levels of pragmatic tolerance in both language groups as participants in both groups rejected underinformative *some* statements most of the time, and critically for RQ2, there was no significant between-group difference in mean answer score to underinformative some statements. Similarly, a categorical analysis found no evidence of a difference between the two language groups in terms of the number of participants categorised as having low levels of pragmatic tolerance. These findings are consistent with those of Dupuy et al. (2019) and Snape and Hosoi (2018) who also did not find between-group differences in binary tasks either (although they framed their interpretation of the results in terms of a theoretical claim about pragmatic competence). However, in contrast to the study at hand, Slabakova (2010) and Mazzaggio et al. (2021) reported significant between-group differences in binary tasks. The discrepancy regarding the binary data between the current study and Slabakova (2010) and Mazzaggio et al. (2021) is due to the fact that the current study revealed null-results (i.e. no between-group differences in a relatively small sample). Nevertheless, following other studies (e.g. Dupuy et al., 2019), I will discuss possible reasons that there might/might not be a difference between groups on this type of binary measure data, whilst bearing in mind the need for caution.

In this vein, previous work has considered several reasons as to why different language groups might react differently (or not) in binary tasks when confronted with scalar expressions (if fundamental between-group differences concerning general cognitive abilities can be ruled out). 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).

In addition to the influence of politeness-contexts and individuals' honesty-traits on the interpretation of scalar expressions, other factors such as social context have been found to influence answer behaviours as well. For example, investigating

pragmatic tolerance, Sikos et al. (2019) manipulated social attributes of (imaginary) speakers and found that a speaker's likeability influenced acceptance rates of his/her underinformative utterances in a binary SJT. If speakers were perceived as likeable, participants were more likely to accept underinformative utterances, and, vice versa, if speakers were perceived less likeable, participants were less likely to accept underinformative items. Interestingly, made-up 'non-native speakers' were rated lowest by the participants (who were native speakers) in terms of likeability, and underinformative utterances of those made-up 'non-native speakers' were most likely to be penalized. Essentially, the social attribute 'non-native speaker' resulted in the highest rejection rates of underinformative utterances. Findings like these as well as Feeney and Bonnefon's (2012) results are important to consider in L2 pragmatics research because they are strong reminders that derivation processes such as implicature derivation are not merely derivational processes but also fluctuant social endeavours influenced by individual differences and social contexts (Gibbs & Colston, 2020).

Individual differences and social contexts could have also been influential in the present study and provide explanations as to why *no* between-group differences were found concerning pragmatic tolerance. The simple fact that social and contextual differences between the British and the German students are marginal might explain the results. For example, both groups were of the same age and both schools were situated in similar socio-economic environments (a grammar school in South-Germany and a private school on the Isle of Wight; see 3.2.1). Regarding honesty-traits, there is no obvious reason to assume major differences between the German and British students. Additionally, politeness played no noticeable role in the experimental input. Moreover, the speaker (the animated chef in the kitchen) was the same for both groups, therefore, the speaker's social attributes did not change between groups and had probably no influence on the results. Combined with the assumption that there are no fundamental cognitive differences between the two language groups (cf. Kecskes, 2019), all these factors could be part of an explanation as to why there were *no* differences detected between groups in terms of pragmatic tolerance in the current study.

Following the same line of argument, social context could account for differences between groups found in Slabakova's (2010) second experiment (see 2.3.1). She found that if context is provided the L2 learners gave significantly more pragmatic answers (*some but not all*) to underinformative *some* statements than the English native speakers. One reason for this difference could lie in the visual input itself. Slabakova (2010) employed picture stories depicting a real girl. One example scenario depicts her eating candies. By the end of the story, the girl has eaten all of the candies. The last picture shows a woman with a look of reproach standing next to the girl, asking her what happened with the candy. The girl then answers *I have eaten some of the candies*. Such situations are not value-free, they carry a social load in several respects, for example, the girl could be the woman's daughter and is thus lying to her mother. It is an open question whether Slabakova's (2010) results are influenced by social characteristics the participants (unconsciously) attributed to the girl, such as deceptiveness or dishonesty. Perhaps, in a socially neutral context, they would have reacted differently. Essentially, corresponding to Sikos et al.'s (2019) findings, it is likely that the picture stories' social contexts influenced participants' behaviour. Interestingly, Slabakova (2010) included a Korean L1 control group that was tested in Korean. This group actually rejected underinformative *some* statements (provided visual context) more often (75% pragmatic answers) than the English L1 group (62,5% pragmatic answers). Although Slabakova (2010) did not compare these two groups statistically in this respect, the direction of the means is in line with this explanation, suggesting that the differences in answer behaviour between the Korean L2 learners of English and the English L1 speakers in this situation with visual context might indeed not *only* be influenced by factors related to L1 or L2, but instead by factors related to broader cultural characteristics. However, note that Slabakova (2010) reported significant between-group differences between the Korean L1 speakers and the Korean L2 learners of English, suggesting that for example cognitive L2 constraints, as discussed in 2.3, might be at play here as well. Nonetheless, perhaps both the English and Korean participants (L1 Korean and L2 English) based their answer decisions in part on cultural and social factors such as attitudes towards dishonesty or deception – irrespective of the test-language. In general, research in the realms of metalinguistic awareness, metalinguistic attitudes and tolerance of pragmatic violations appears to be

partially influenced by non-linguistic factors. Such factors should be considered when evaluating binary SJT data.

In contrast, social- and context-factors are presumably negligible in Mazzaggio et al.'s (2021) study. Unlike Slabakova (2010), they report that L2 learners gave *fewer* pragmatic answers than native speakers, however, they provided no visual context and solely built on participants' encyclopaedic knowledge (e.g. *Some elephants have trunks*). Instead, they deliberately increased the participants' cognitive burden through time limits and by presenting the input in aural form. Potentially, these cognitive obstacles influenced their L2 learners' behaviour, because, as discussed in 2.3, L2 learners even at advanced stages have slight cognitive disadvantages compared to native speakers. Perhaps because of time pressure and the experiments' aural characteristic, the L2 participants had fewer cognitive resources available to access their entire repertoire of metalinguistic awareness and attitudes, and thus less opportunity to examine statements relative to these metalinguistic properties. This could have led them to (automatically) accept underinformative *some* statements more often than native speakers. In comparison, L1 speakers could access their full repertoire of metalinguistic awareness and attitudes and, therefore, make informed decisions and show lower levels of pragmatic tolerance.

5.3 Pragmatically oblivious and inconsistent behaviour

On the level of individual participants' behaviour, note that the quinary measure data indicated that three participants in the English L1 group showed no sensitivity to underinformativeness at all because they always agreed with underinformative *some* statements. Moreover, some participants in both language groups did not answer consistently which indicates that they too did not show sensitivity to underinformativeness in at least one trial (see Table 3 in 4.1.2). Similarly, the binary measure data indicated that some participants in both language groups showed inconsistent answer behaviours in terms of their pragmatic tolerance (see Table 4 and Table 5 in 4.2.2). Contrary to these phenomena, one might expect participants to answer consistently, for example, to be sensitive to underinformativeness in *each* trial (in the quinary measure) or to be pragmatically tolerant in *each* trial (in the binary measure). This raises the questions (a) why some participants were apparently

pragmatically oblivious in the quinary measure, (b) why some participants answered inconsistently in both the quinary and binary measure, and (c) what triggered this behaviour each time.[7]

Regarding the apparent pragmatic obliviousness, it seems unlikely that the three English L1 participants who always agreed with underinformative *some* statements are indeed pragmatically oblivious since they are competent native speakers with no learning impairments. However, the overall observation that 50% of all answers to underinformative *some* items in the English L1 group were 'agree' answers is striking. Interestingly, such non-rejections of underinformative *some* items by competent adult native speaker controls have been reported repeatedly (e.g. Guasti et al., 2005; Noveck, 2001; Papafragou & Musolino, 2003), however, previous L2 studies did not report their data at the level of individual participants. Concerning this phenomenon, Katsos and Smith (2010) point out that from a theoretical pragmatics perspective these participants do nothing 'wrong' because, logically speaking, underinformative *some* statements are in fact correct (see 2.1). Against the background of pragmatic tolerance, they argue that categorical rejections of underinformative *some* items do not solely rely on someone's sensitivity to underinformativeness but also on someone's disposition towards pragmatic violations. Recall that the statistical analysis of the quinary data from the current study suggested that there is evidence that both language groups were sensitive to underinformativeness. Therefore, conforming to Katsos and Smith (2010), it seems plausible that both the (apparently) pragmatically oblivious behaviour by three English L1 speakers and the inconsistent answer behaviours by some participants in both language groups in both the quinary and the binary measure are results of somewhat competing decision-making processes, namely, the *detection* of pragmatic violation on the one hand and the *tolerance* of pragmatic violation on the other hand. This is supported by the fact that the inconsistent participants in the quinary measure demonstrated both sensitivity to underinformativeness *and* apparent pragmatic obliviousness throughout the task. That means that each time they encountered an underinformative *some* item they balanced which characteristic to privilege. In fact, there seems no reason to assume that pragmatic tolerance mechanisms (which were

[7] *Pragmatically oblivious* and *pragmatic obliviousness* seem to be somewhat inapplicable terms; however, they are used in the relevant literature (e.g. Slabakova, 2010).

captured by the binary measure) should play no role in the quinary measure. On the contrary, it is important to acknowledge that participants potentially make metalinguistic judgements all the time, in both measure types, especially when there is no time limit. By means of binary/graded SJTs there is no possibility to disentangle the potentially intertwined 'detection vs. tolerance of pragmatic violation' reasoning processes. Therefore, regarding the question of what triggered participants' reaction in one way or another in different trials during the same task, the data at hand is insufficient. For future work, there are possible manipulations which could address this: For example, one could manipulate the social attributes of the (animated) speaker to determine participants' reasoning regarding their metalinguistic judgements. It could also be useful to include participant characteristics such as age, gender or cultural background in the statistical analysis and examine relevant statistical associations. Thus, future research could determine which context cues or participant characteristics tip the balance with regards to the competition between sensitivity to underinformativeness and pragmatic tolerance.

5.4 Limitations

It is important to outline some methodological limitations to the current study which may have influenced findings and limit the generalizability of the outcomes.

Native language testing. Unlike previous studies such as Slabakova (2010), the current study included no German L1 control group (tested in German) to rule out culture-specific pragmatic biases. A German L1 group would have been critical if we had seen significant differences between the German EFL and the English L1 group, for example regarding their pragmatic tolerance. If so, the German native speaker group would have allowed us to ask whether the differences were due to testing in an L1 versus an L2, or actually because participants are Germans or Britons (i.e. cultural differences). However, as there was no evidence for between-group difference in both measure types, the German L1 group was not necessary. Nonetheless, in the context of research into pragmatic tolerance where social and cultural factors seem to play a role, L1 control groups might be important for future studies which, perhaps, use participants that come from more diverse and culturally more distant backgrounds.

Sample sizes. The a-priori power analyses reported in 3.2 indicated that larger sample sizes would have been required to assure there is large enough power to find between-group differences if there are any. This is particularly important in the context of the current study because it presents inferences about null results; this must be treated cautiously given the low power. In future, a high-powered replication of the current study would be interesting – ideally one that incorporated other statistics such as the Bayes Factor (cf. Dienes, 2008) or equivalence testing (cf. Goertzen & Cribbie, 2010; Quertemont, 2011), which allow evaluations of evidence of H0 as well as H1. However, the researcher decided that although participant recruitment via schools reduced overall sample sizes compared to the possibilities of online recruitment, the homogeneity of samples provided in schools and the guarantee of recruiting L2 learners with a standardized level of language proficiency within the time constraints, outweighed the fact that the study is underpowered given practicalities. A lack of power might have been an issue in previous studies too, for example in Snape and Hosoi (2018). They found no significant between-group differences (using a binary SJT), although the experimental set-up was similar to other studies (e.g. Slabakova, 2010). Interestingly, their sample sizes per group were only half the size (n/group ≈ 15) of those of other studies that found significant differences, such as Slabakova (2010) (n/group ≈ 30). Therefore, replication studies should recruit more participants per language group to improve reliability and generalizability.

Effect of experimental situation. It is widely acknowledged in experimental pragmatics research that the experimental situations themselves influence participants' behaviour (Gibbs & Colston, 2020) – especially when the investigations examine participants' metalinguistic judgements. Regarding the current study, participants were asked to answer quickly (i.e. to get their most intuitive answer), however, as for example demonstrated above with the German EFL quinary answer behaviour to optimally false *all* items, participants might have learned to give more nuanced answers throughout the experiment. Therefore, it generally cannot be ruled out that participants (a) figured out what the tasks were about and (b) subsequently behaved in ways they thought they would be expected to (including the underinformative items).

Lack of qualitative data. The binary measure revealed low levels of pragmatic tolerance in both language groups. It would have been interesting to investigate why

pragmatic tolerance levels were generally low in both groups, however, the main research aim of the current study was to examine between-group differences regarding pragmatic tolerance. Besides, the quantitative data gave no possibility to investigate reasons for the participants' behaviour further. Nonetheless, qualitative data, for example from post-task interviews, would have been an interesting instrument to investigate participants' answer behaviour (i.e. their metalinguistic attitudes) in more detail and to examine why pragmatic tolerance levels were generally low.

6 Conclusions and Implications

Despite the aforementioned limitations, the current study's findings are consistent with previous research in other domains of Applied Linguistics research demonstrating that graded judgement tasks provide more nuanced insights into participants' pragmatic abilities. Essentially, the findings demonstrate that German EFL learners are sensitive to underinformativeness, that means they are pragmatically competent in this respect, and there was no evidence that they differed from English L1 speakers in this regard. Therefore, the current study expands methodological criticisms regarding the use of binary judgement tasks in scalar implicature interpretation research to the domain of L2 research and provides empirical research support for such methodological criticisms in the domain of L2 research. Additionally, regarding pragmatic tolerance, the findings overall demonstrate low levels of pragmatic tolerance, and with again no evidence that the German EFL learners differed in their pragmatic tolerance compared with the English L1 speakers, nor that there was a difference between the language groups in terms of the proportion of participants categorized as having/not having low levels of pragmatic tolerance.

In light of such findings, the current study contributes to L2 research by highlighting the important distinction between an individual's pragmatic competence and his/her pragmatic tolerance. It is crucial in L2 research to acknowledge that binary judgement tasks might not be the appropriate instrument to draw conclusions about L2 learners' pragmatic competence. The change from binary to graded measures is a methodological advance because it prevents categorical misconceptions about L2 learners' pragmatic competence. In fact, a rethinking concerning experimental methods will allow more nuanced and informative insights into L2 learners' pragmatic abilities. A similar process in L1 research has led to changes in the way children's pragmatic abilities are perceived and, meanwhile, acknowledged.

As the current study represents basic research it has no direct implications on language teaching. However, it contributes to L2 teaching indirectly because it acknowledges L2 learners' pragmatic competence, and it introduces the concept of pragmatic tolerance into the realm of L2 learning. On the one hand, previous research might have seemed to suggest that there were pragmatic competence differences

between L1 speakers and L2 learners, and that this issue might be targeted in teaching. However, the current study's findings are consistent with an account in which L2 learners make inferences in a similar way to L1 speakers, that is, inference making is a cognitive skill that transfers without needing to be taught, provided the context conditions are right in terms of, for example, visual support – although this awaits confirmation from high-powered replication studies (see 5.4). On the other hand, the concept of pragmatic tolerance is promising for L2 teaching as other studies (e.g. Feeney & Bonnefon, 2012) found that metalinguistic attitudes and social factors influence participants' pragmatic tolerance. A general awareness for pragmatic tolerance in L2 teaching and learning and an acknowledgment of influential factors such as a speaker's social characteristics (cf. Sikos et al., 2018) might be beneficial for learners regarding their understanding of intercultural communication and their behaviour towards challenges in intercultural communication, such as misunderstandings. While the current study found no evidence of between-group differences regarding pragmatic tolerance, future studies with participant groups from less similar cultural backgrounds might find differences and, thus, raise important issues for dealing with metalinguistic awareness in classrooms.

From a theoretical perspective, a replication of the current study which finds evidence for the null hypotheses would contribute to L2 linguistic theory, for example, concerning the Interface Hypothesis (Sorace, 2011, 2012; Sorace & Filiaci, 2006) which claims that linguistic phenomena at the interface of linguistic components (e.g. syntax) and pragmatic/discourse components remain challenging even for highly proficient L2 learners. Such types of interfaces are referred to as 'external' interfaces because they rely on associations between the grammatical system and non-linguistic context (Slabakova, 2015). Although the theory is not uncontested (e.g. Lardier, 2011; White 2011), its core idea remains: unlike other L2 phenomena, external interfaces cannot be acquired, even at advanced stages (Destruel & Donaldson, 2017). Essentially, the hypothesis predicts that at external interfaces even advanced L2 learners will make recurring mistakes such as L1 transfer or "reliance on pragmatically neutralized default forms" (Destruel & Donaldson, 2017; cf. Sorace, 2011). While findings of some of the previous L2 scalar implicature studies which found significant between-group differences (e.g. Mazzaggio et al., 2021; Slabakova, 2010) seem to

provide evidence for the Interface Hypothesis (the L2 participants in these studies were reasonably proficient), if future studies can establish evidence for patterns-of-no-difference (like in the current study), this will indicate that in fact German EFL learners do not perform worse than English L1 speakers at the syntax-pragmatics interface of scalar implicature interpretation. This would speak against the hypothesis that in terms of ultimate attainment proficient German EFL learners should perform worse than English L1 speakers in a situation where non-linguistic modules are involved (i.e. context-considerations).

With regards to future research, regardless of whether pragmatic competence or tolerance is investigated, the present study's power analyses showed that larger sample sizes are required. Once the current patterns of results are established in high-powered replications with somewhat different statistics, future research should diversify its methodological approach to gain more precise insights into scalar implicature derivation processes while reducing the effects of metalinguistic reasoning on the results, for example, by including eye- and/or mouse-tracking paradigms. That way, conscious reasoning processes can be controlled and ruled out to at least some extend. For example, L1 studies investigating the processing time-course of scalar implicatures such as Huang and Snedeker's (2009) visual-world paradigm study measured participants' eye movements within the range of milliseconds after hearing underinformative input sentences. Measurements in such small time-ranges minimize the possibility that participants consciously think about the interpretation, instead reflecting participants' automatized processing behaviour. In a similar L2 context, the effects of context provision on the time-course and stages of scalar implicature derivation could be investigated. Moreover, the potential effects of proficiency on scalar implicature derivation abilities in an L2 could inform pragmatic theory.

Additionally, research concerning L2 learners' pragmatic tolerance could examine why pragmatic tolerance levels are generally high/low in certain language groups, how L2 learners outbalance the *pragmatic competence – pragmatic tolerance* dissonance compared to native speakers and which immediate context cues or participant characteristics influence this process. Quantitative research could be complemented by qualitative methods, for example, via post-task interviews. Moreover, in collaboration with sociolinguistics, effects of wider linguistic and cultural attitudes and backgrounds

as well as effects of different target languages on pragmatic tolerance are promising in L2 research. In this context, an expansion of pragmatic tolerance inquiry to other types of implicatures such as quality implicatures and to other pragmatic phenomena such as presuppositions is auspicious because findings are likely to contribute to our understanding of cross-cultural communicative misunderstandings and linguistic stereotyping (Thomas, 1983; Padilla Cruz, 2018).

7 References

Adesope, O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research* 80(2), 207-245.

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.

Belletti, A., Bennati, E., & Sorace, A. (2007). Theoretical and developmental issues in the syntax of subjects: Evidence from near-native Italian. *Natural Language & Linguistic Theory* 25, 657-689.

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

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

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.

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

Breheny, R. (2019). Scalar Implicatures. In C. Cummins & N. Katsos (Eds.). *The Oxford Handbook of Experimental Semantics and Pragmatics* (pp. 39-62). Oxford: Oxford University Press.

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

Chemla, E. & Singh, R. (2014a). Remarks on the Experimental Turn in the Study of Scalar Implicature, Part I. *Language and Linguistics Compass* 8(9), 373-386.

Chemla, E. & Singh, R. (2014b). Remarks on the Experimental Turn in the Study of Scalar Implicature, Part II. *Language and Linguistics Compass* 8(9), 387-399.

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

Council of Europe. (2021, August). *Global scale - Table 1 (CEFR 3.3): Common Reference levels.* https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale

Cummins. C. (2019). *Pragmatics*. Edinburgh: Edinburgh University Press.

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

Degen, J. & Tanenhaus, M. K. (2011). Making inferences: the case of scalar implicature processing. *Proceedings of the Cognitive Science Society* 33(33), 3299-3304.

Degen, J. & Tanenhaus, M. (2015). Processing Scalar Implicature: A Constraint-Based Approach. *Cognitive Science* 39, 667-710.

Degen, J. & Tanenhaus, M. (2016). Availability of Alternatives and the Processing of Scalar Implicatures: A Visual World Eye-Tracking Study. *Cognitive Science* 40, 172-201.

Destruel, E. & Donaldson, B. (2017). Second language acquisition of pragmatic inferences: Evidence from the French c'est-cleft. *Applied Psycholinguistics* 38, 703-732.

Dienes, Z. (2008) *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. New York: Palgrave Macmillan.

Dörnyei, Z. (2007). *Research methods in applied linguistics*. New York: Oxford University Press.

Doyle, John J. (1951). In defense of the square of opposition. *The New Scholasticism* 25(4), 367-96.

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.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behaviour Research Methods* 41, 1149-1160.

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.

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). London: Sage Publications.

Gazdar, G. (1979). *Pragmatics: Implicature, Presupposition and Logical Form*. New York: Academic Press.

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

Gibbs Jr., R. W. & Colston, H. L. (2020). Pragmatics Always Matters: An Expanded Vision of Experimental Pragmatics. *Frontiers in Psychology* 11: 1619.

Goertzen, J. R. & Cribbie, R. A. (2010). Detecting a lack of association: An equivalence testing approach. *British Journal of Mathematical and Statistical Psychology* 63, 527-537.

Grassmann, S., Stracke, M., & Tomasello, M. (2009). Two-year-olds exclude novel objects as potential referents of novel words based on pragmatics. *Cognition* 112(3), 488-493.

Green, G. M. (1990) The Universality of Gricean Interpretation. *The Annual Proceedings of the Berkeley Linguistics Society* 16(1), 411-428.

Grice, H. P. (1975) Logic and Conversation. In P. Cole & H. Morgan (Eds.). *Syntax and Semantics. Vol. 3: Speech Acts* (pp. 41-58). New York: Academic Press.

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

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

Gürel, A., (2006). L2 acquisition of pragmatic and syntactic constraints in the use of overt and null subject pronouns. In R. Slabakova, S. Montrul, & P. Prévost (Eds.). *Inquiries in Linguistic Development* (pp. 259-282). Amsterdam: John Benjamins.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics.* New York, NY: Newbury House.

Haugh, M. (2013) Inference and Implicature. In C. A. Chapelle (Ed.). *The Encyclopedia of Applied Linguistics. Volume V* (pp. 2658-2665). Oxford: Blackwell Publishing.

Haugh, M. (2015). *Im/Politeness Impliactures*. Berlin: de Gruyter Mouton.

Holtgraves, T., Kwon, G., & Zelaya, T. M. (2019). Psycholinguistic Approaches to L2 Pragmatics Research. In N. Taguchi (Ed.). *The Routledge Handbook of Second Language Acquisition and Pragmatics* (pp. 272-284) (1st ed.). New York: Routledge.

Horn, L. R. (1972). *On the Semantic Properties of Logical Operators in English*. PhD dissertation, UCLA. Distributed by the Indiana University Linguistics Club, 1976.

Horn, L. R. (2006). Implicature. In L. R. Horn & G. Ward (Eds.). *The Handbook of Pragmatics* (pp. 2-28). Oxford: Blackwell Publishing.

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.

Kallia, A. (2007). *Politeness and Implicatures. Expanding the Cooperative Principle*. Hamburg: Verlag Dr. Kovač.

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

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 Development – Proceedings* (pp. 221-232). Boston: Cascadilla Press.

Kecskes, I. (2019). *English as a Lingua Franca: The Pragmatic Perspective*. Cambridge: Cambridge University Press.

Kultusministerium Baden-Württemberg. (2016). Bildungspläne Baden-Württemberg. Gymnasium – Englisch als erste Fremdsprache. Retrieved from: https://www.bildungsplaene-bw.de/,Lde/LS/BP2016BW/ALLG/GYM/E1

Kutas, M., & Hillyard, S. A. (1980a). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science* 207, 203-205.

Kutas, M., & Hillyard, S. A. (1980b). Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307, 161-163.

Lardiere, D. (2011). Who is the Interface Hypothesis about? *Linguistic Approaches to Bilingualism* 1(1), 48-53.

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

Liedtke, F. (2016) *Moderne Pragmatik.* Tübingen: Narr Francke Attempto.

Lin, Y. (2016). Processing of Scalar Inferences by Mandarin Learners of English: An Online-Measure. *PLoS ONE* 11(1).

Lozano, C. (2006). Focus and split intransitivity: the acquisition of word order alternations in non-native Spanish. *Second Language Research* 22, 145-187.

Magri, G. (2009). Mismatching scalar implicatures. In P. Égré and G. Magri (Eds.). *Presuppositions and Implicatures: Proceedings of the MIT-Paris Workshop* (pp. 153-168). Cambridge, MA: MIT Working Papers in Linguistics.

Magri, G. (2011). Another argument for embedded scalar implicatures based on oddness in downward entailing contexts. *Proceedings of SALT* 20, 564-581.

Mazzaggio, G., Panizza, D., & Surian, L. (2021). On the Interpretation of Scalar Implicatures in First and Second Language. *Journal of Pragmatics* 171, 62-75.

Muijs, D. (2011). *Doing Quantitative Research in Education with SPSS*. London: SAGE Publications.

Ninio, A., & Snow, C. (1996). *Pragmatic Development*. Boulder, CO: Westview Press.

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

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

Padilla Cruz, M. (2018) Pragmatic Competence Injustice. *Social Epistemology* 32 (3), 143-163

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

Pipijn, K. & Schaeken, W. (2012). Children and Pragmatic Implicatures: A Test of the Pragmatic Tolerance Hypothesis with Different Tasks. *Proceedings of the Annual Meeting of the Cognitive Science Society* 34, 2186-2191.

Quertemont, E. (2011). How to statistically show the absence of an effect. *Psychologica Belgica* 51, 109-127.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (2nd ed.). Newbury Park, CA: SAGE Publications.

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). Amsterdam: John Benjamins.

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.

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

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

Slabakova, R. (2015). The effect of construction frequency and native transfer on second language knowledge of the syntax–discourse interface. *Applied Psycholinguistics* 36, 671-699.

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.

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

Sorace, A. (2012). Pinning down the concept of interface in bilingual development: A reply to peer commentaries. *Linguistic Approaches to Bilingualism* 2, 209-216.

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

Southgate, V., Chevallier, C. & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science* 13(6), 907-912.

Sperber, D. & Wilson, D. (1986). *Relevance: Communication and Cognition* (2nd ed.) Cambridge, MA: Harvard University Press.

Sperber, D. & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind & Language* 17, 3-23.

Storto, G., Tanenhaus, M. (2005). Are scalar implicatures computed online? In: E. Maier, C. Bary, & J. Huitink (Eds.). *Proceedings of Sinn und Bedeutung* 9 (pp. 431-445). Nijmegen: Centre for Semantics.

Thomas, J. (1983) Cross-Cultural Pragmatic Failure. *Applied Linguistics* 4 (2), 91-111.

Tomlinson, J. M., Bailey, T. M., & Bott, L. (2013). Possibly all of that and then some: scalar implicatures are understood in two steps. *Journal of Memory and Language* 69(1), 18-35.

University of Oxford Central University Research Ethics Committee. (2020). *Best Practice Guidance 04_Version 2.3 for Research Involving Competent Youths*. Retrieved from: https://researchsupport.admin.ox.ac.uk/files/bpg04competentyouthspdf-0.

Veenstra, A., Hollebrandse, B., & Katsos, N. (2017). Why some children accept under-informative utterances. Lack of competence or Pragmatic Tolerance? *Pragmatics & Cognition* 24(2), 297-314.

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). Berlin/Boston: de Gruyter Mouton.

Waldon, B. & Degen, J. (2020). Modeling Behavior in Truth Value Judgment Task Experiments. *Proceedings of the Society for Computation in Linguistics* 3.

White, L. (2011). The interface hypothesis: How far does it extend? *Linguistic Approaches to Bilingualism* 1, 108-110.

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). Austin, TX: Cognitive Science Society.

8 Appendix

Appendix 1 – Power Analyses

Appendix 1a – Power Analysis RQ1

The power analysis was conducted using G*Power Version 3.1.

Power Analysis for Research Question 1 – quinary data:

Statistical Test:	Wilcoxon signed-rank test (within-group)
Tails:	2
Effect size:	.7
Alpha-error probability:	.05
Level of power:	.8

Alpha-error probability and level of power are adopted from Field (2018: 139).
The effect size was adopted from a previous study with similar study design and research aim (Katsos & Bishop, 2011).

Calculation result:

Total sample size (per language group):	19 [actual n in the current study is 19 German EFL and 16 English L1]
Actual power:	.8

Appendix 1b – Power Analysis RQ2

The power analysis was conducted using G*Power Version 3.1.

Power Analysis for Research Question 2 – binary data:

Statistical Test:	Mann-Whitney U-test (two groups)
Tails:	2
Effect size:	.7
Alpha-error probability:	.05
Level of power:	.8

Alpha-error probability and level of power are adopted from Field (2018: 139).

The effect size was adopted from a previous study with similar study design and research aim (Katsos & Bishop, 2011).

Calculation result:

Total sample size (per language group):	35 [actual n in the current study is 17 German EFL and 14 English L1]
Actual power:	.8

Appendix 2 – Example Stimuli

Note: The blue bars in the experimental items are the progress bars.

Appendix 2a – quinary measure

Fig. A9 Example of an experimental item quinary measure (optimally true statement)

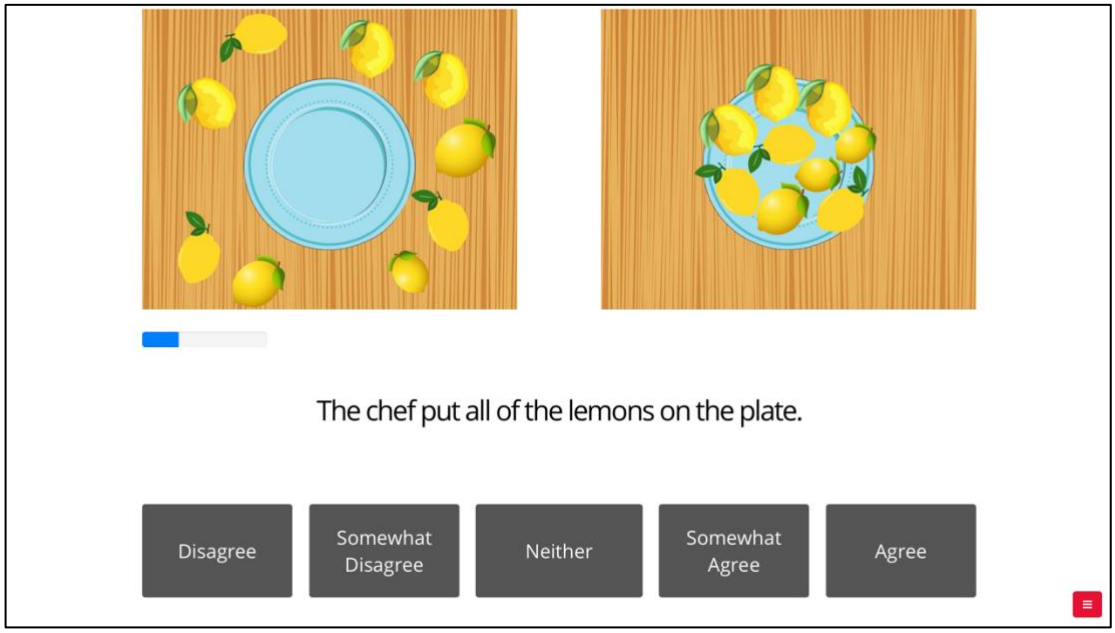


Fig. A10 Example of an experimental item quinary measure (optimally true statement)

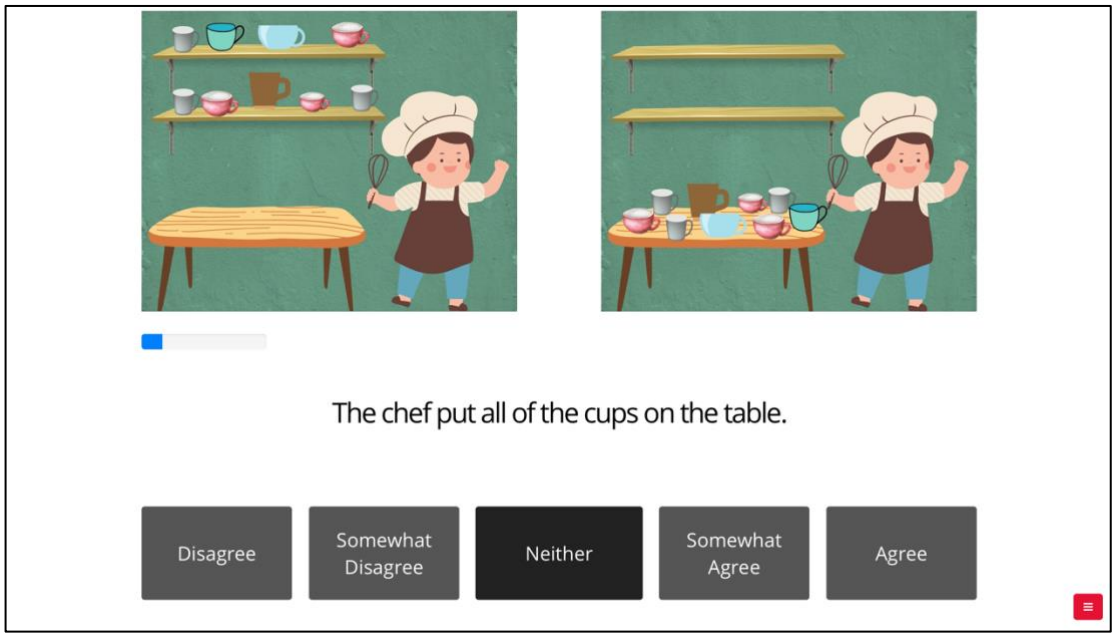
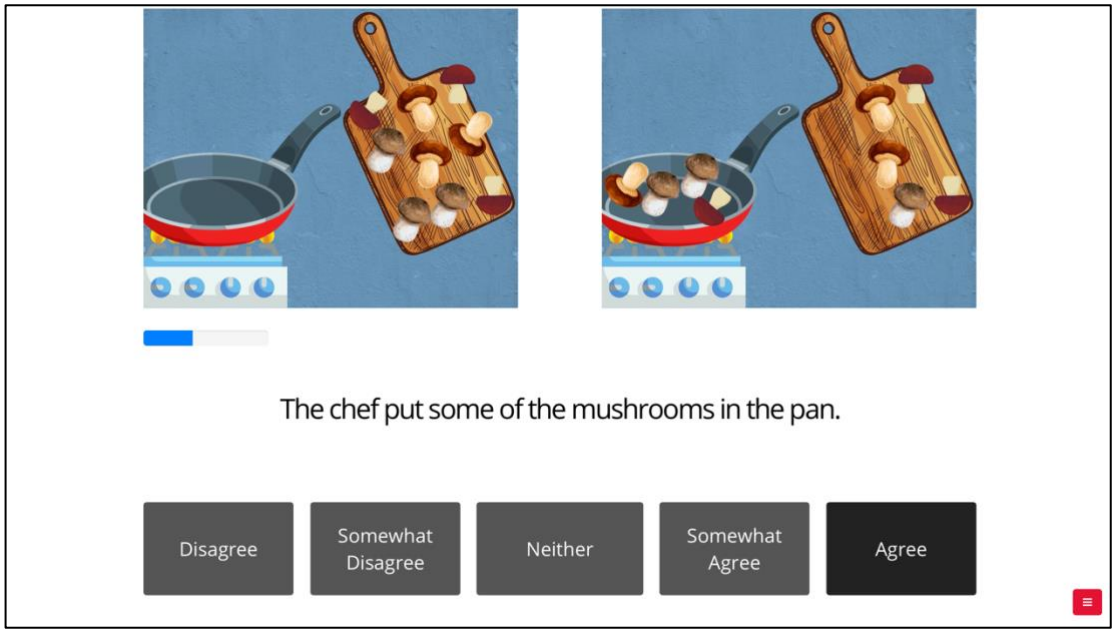


Fig. A11 Example of an experimental item quinary measure (felicitous some statement)



Appendix 2b – binary measure

Fig. A12 Example of an experimental item binary measure (underinformative some statement)

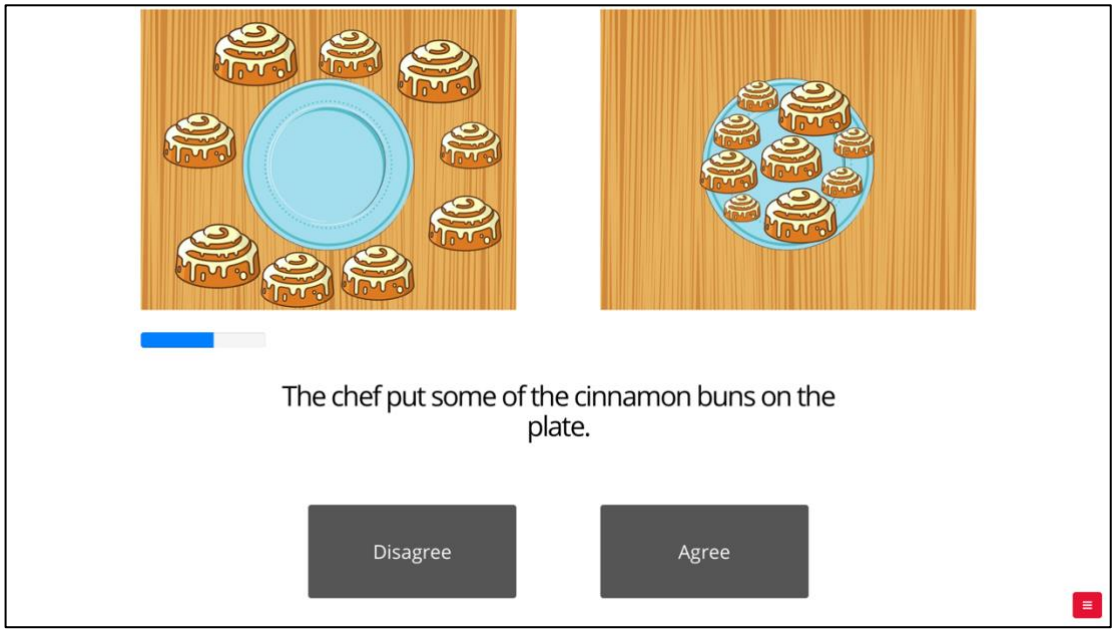


Fig. A13 Example of an experimental item binary measure (optimally false all statement)

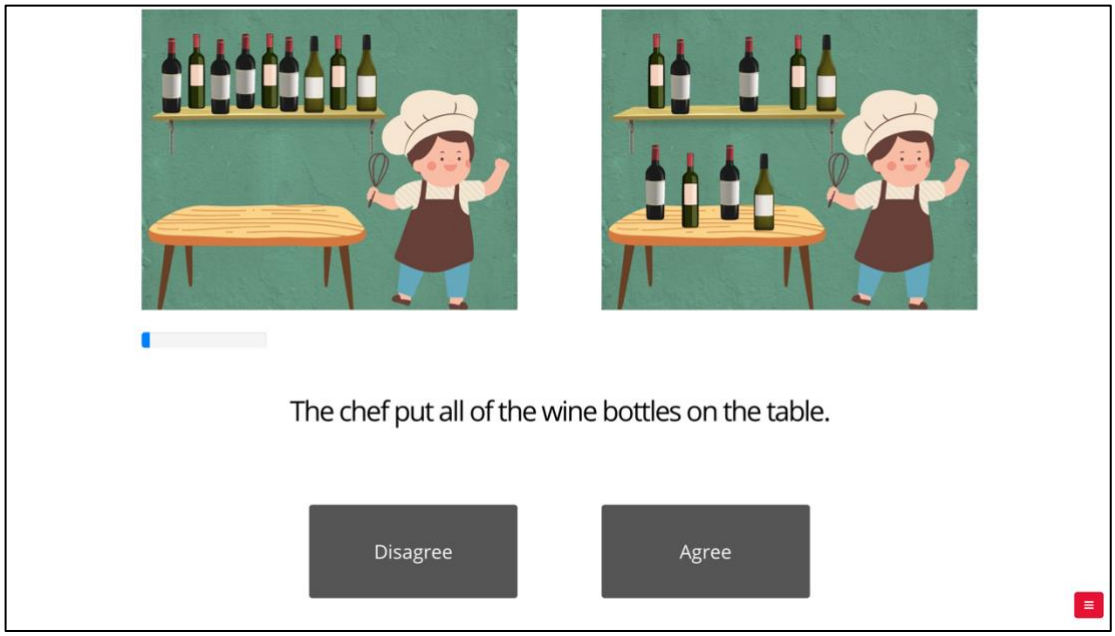
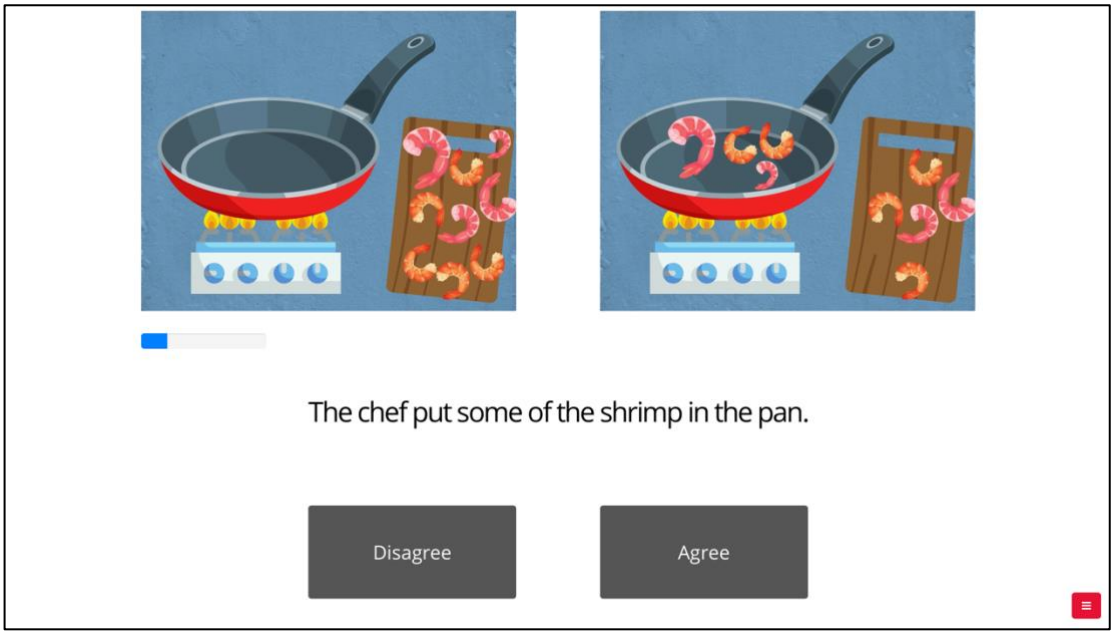


Fig. A14 Example of an experimental item binary measure (felicitous some statement)



Appendix 3 – Online Participant Information Sheet

Title of Study: Interpretation of scalar terms by native speakers and foreign language learners

This research has been approved by the Department of Education Research Ethics Committee (DREC) Ref. No.: ED-CIA-21-164.

Department:
Department of Education
University of Oxford
15 Norham Gardens
Oxford
OX2 6PY

Name and Contact Details of the Researcher:
Johannes Schulz
Master of Science student
johannes.schulz@education.ox.ac.uk

Name and Contact Details of the Principal Investigator:
Prof Dr Elizabeth Wonnacott
Associate Professor in Applied Linguistics
elizabeth.wonnacott@education.ox.ac.uk

1. **Invitation Paragraph**
   Dear participant,

   You are being invited to take part in a research project investigating how humans process language. Before you decide, it is extremely important for you to **READ CAREFULLY** in order to understand why the research is being done and what participation will involve. We wish to remind you that your participation is completely voluntary and you have the right to opt-out at any time without explanations.

   Please read the following document, download it if you want, and take time to decide whether or not you wish to take part in our research. If there is anything that is not clear or if you wish to give us feedback before or after your participation, do not hesitate to contact us (see above our emails).

   Thank you for reading this and for considering to take part in our experiment.

2. **What is the project's purpose?**
   In this study we are interested in understanding how human process language in their mind. In order to do so, this experiment includes tasks which present you with a sentence and requires you to evaluate whether you 'Agree' or 'Disagree' with it

as fast and accurate as possible. These tasks would feel very easy (and perhaps pointless) to you, but are invaluable tools for us to understand specific domains of language processing. In the end of the experiment, we will debrief you more in detail of what specifically this project is about.

Bear in mind that this experiment doesn't measure *your* evaluation or attentional abilities, but rather a general behavioural tendency of different groups of communicators. For this reason, if you do agree to participate in this study, we wish to ask you to do your best without thinking that this is some sort of a "test". This experiment doesn't have any clinical relevance, and doesn't tell us anything about *your* intellectual capabilities or *your* personality. There are no *right* or *wrong* answers.

3. Why have I been chosen?

You have been selected because:

Inclusion criteria:

1. You are a native speaker of English **or** you are a native speaker of German learning English as a foreign language
2. You are a student at either a German grammar school **or** at a British senior school **or** you have received an invitation from the researcher via social media
3. You are at least 16 years old or older
4. You declare that you don't have any neurological and learning impairments
5. You have normal or corrected-to-normal vision

Exclusion criteria:

Unfortunately, you cannot take part if:

1. You are a native speaker of a language other than German or English.
2. You have been raised bilingually (i.e. at your home, two or more languages have been used throughout your childhood)
3. You are younger than 16 years
4. You have any neurological and learning impairments
5. You do not have normal or corrected to normal vision (i.e. wearing glasses is fine!)

4. Do I have to take part?

No. It's up to you to decide whether or not to take part—your participation is entirely voluntary.

Your refusal to agree won't result in any penalties or prejudice. However, if you do agree in taking part to our experiment, you'll be asked to declare your decision digitally by completing a consent form (in the next page). Remember that you can always opt-out at any time without giving explanations, and if you do decide to withdraw from the study you'll be asked what you wish to happen to the data you have provided to that point.

5. **What will happen to me if I take part?**
   If you do decide to take part in this study, you'll be presented with individual short statements, one after another. Some statement may be accompanied by a picture. Your task is to simply read the statements, look at the pictures, and evaluate as fast and accurate as possible whether you agree with the statement or disagree by clicking keys on your keyboard (or buttons on your touchscreen). A three-second countdown will be presented in the corner of the screen. You have to make a decision within this timeframe.

   Before the experiment, you'll be asked to familiarise with the task.

   If you decide to participate, you would be busy for 15 min max in total.

6. **What are the possible disadvantages and risks of taking part?**
   There are no foreseeable discomforts, disadvantages and risks associated to this experiment. However, if you feel or believe there are any unexpected discomforts, disadvantage and risk which arises during the research, you should immediately contact us via email (see below, "what if something goes wrong" section).

7. **What are the possible benefits of taking part?**
   Whilst there are no *direct and immediate* benefit for you in taking part, it is hoped that this study would advance our understanding of the human language processing behaviour and help shaping new educational practices.

8. **What if something goes wrong?**
   If you have any queries or concern about any aspect of this study please contact: Johannes Schulz, Department of Education, University of Oxford, 15 Norham Gardens, Oxford, OX2 6PY **or** Prof Dr Elizabeth Wonnacott, Department of Education, University of Oxford, 15 Norham Gardens, Oxford, OX2 6PY. **Email: johannes.schulz@education.ox.ac.uk** or **elizabeth.wonnacott@education.ox.ac.uk**. If you are unhappy and wish to complain formally, you can do this through the Chair, Department of Education Research Ethics Committee (DREC) Prof Dr Liam Gearon **liam.gearon@education.ox.ac.uk**.

9. **Will my taking part in this project be kept confidential?**
   All the information that we collect about you during the course of the research are related **only** to your performance in this experiment. We **do not** have access to your identity at any point. This information will be kept strictly confidential and you will not be able to be personally identified in any ensuing reports or publications. We make sure that this is the case by assigning you an ID that is generated randomly at the beginning of the experiment. We do not ask your date of birth or any other personal information that might connect you to the ID.

10. **What will happen to the results of the research project?**
    Results will be disseminated via presentation in a Master of Sciences thesis.

If you wish to be informed of the results of the study, do not hesitate to contact us. However, we won't be able to give you feedback about your performance specifically, since we can't identify you, but only give you information at the group level.

11. Local Data Protection Privacy Notice

Notice:
The controller for this project will be the University of Oxford. The University of Oxford Data Protection Officer provides oversight of the university's activities involving the processing of personal data, and can be contacted at data.protection@admin.ox.ac.uk.

This 'local' privacy notice sets out the information that applies to this particular study. Further information on how the University of Oxford uses participant information can be found in our 'general' privacy notice:

For participants in research studies, click here.

The information that is required to be provided to participants under data protection legislation (GDPR and DPA 2018) is provided across both the 'local' and 'general' privacy notices.

12. Who is organising the research?

This project is organized by Johannes Schulz and Prof Dr Elizabeth Wonnacott.
If you have any queries or concern about any aspect of this study please contact: Prof Dr Elizabeth Wonnacott, Department of Education, University of Oxford, elizabeth.wonnacott@education.ox.ac.uk and she will do her best to answer your query within 10 working days to give you an indication of how it will be dealt with. If you remain unhappy or wish to make a formal complaint, please contact the Chair of the Research Ethics Committee at the University of Oxford who will seek to resolve the matter as soon as possible:
Dr Liam Gearon, Departmental Research Ethics Committee (DREC) Chair, Department of Education, 15 Norham Gardens, Oxford, United Kingdom, Email: liam.gearon@education.ox.ac.uk

Contact for further information
You can download this information sheet by clicking on the "download" button. If you do agree to participate in this study, then you can click to "next" at the bottom page and you will be redirected to the Online Consent Form. Before continuing, we wish to remind you that you can contact us for any queries to our emails (Johannes Schulz johannes.schulz@education.ox.ac.uk or Prof Dr Elizabeth Wonnacott elizabeth.wonnacott@education.ox.ac.uk).

Thank you for reading this information sheet and for considering to take part in this research study.

Johannes Schulz & Elizabeth Wonnacott

Appendix 4 – Online Consent Form

Please complete this form after you have read the Participant Information Sheet.

Title of Study: Interpretation of scalar terms by native speakers and foreign language learners

Department: Department of Education

Name and Contact Details of the Researcher: Johannes Schulz, johannes.schulz@education.ox.ac.uk

Name and Contact Details of the Principal Researcher: Prof Dr Elizabeth Wonnacott, elizabeth.wonnacott@education.ox.ac.uk

University of Oxford Data Protection Officer: data.protection@admin.ox.ac.uk

This research has been approved by the Department of Education Research Ethics Committee (DREC) Ref. No.: ED-CIA-21-164.

Thank you for considering taking part in this research.

If you have any queries or concern about any aspect of this study please contact: Johannes Schulz, Department of Education, University of Oxford, 15 Norham Gardens, Oxford, OX2 6PY or Prof Dr Elizabeth Wonnacott, Department of Education, University of Oxford, 15 Norham Gardens, Oxford, OX2 6PY. Email: johannes.schulz@education.ox.ac.uk or elizabeth.wonnacott@education.ox.ac.uk. If you are unhappy and wish to complain formally, you can do this through the Chair, Department of Education Research Ethics Committee (DREC) Prof Dr Liam Gearon Liam.gearon@education.ox.ac.uk. Your school will be given a copy of this Consent Form to keep and refer to at any time.

I confirm that I understand that by ticking each box below I am consenting to this element of the study. I understand that it will be assumed that unticked boxes means that I DO NOT consent to that part of the study. I understand that by not giving consent for any one element that I may be deemed ineligible for the study.

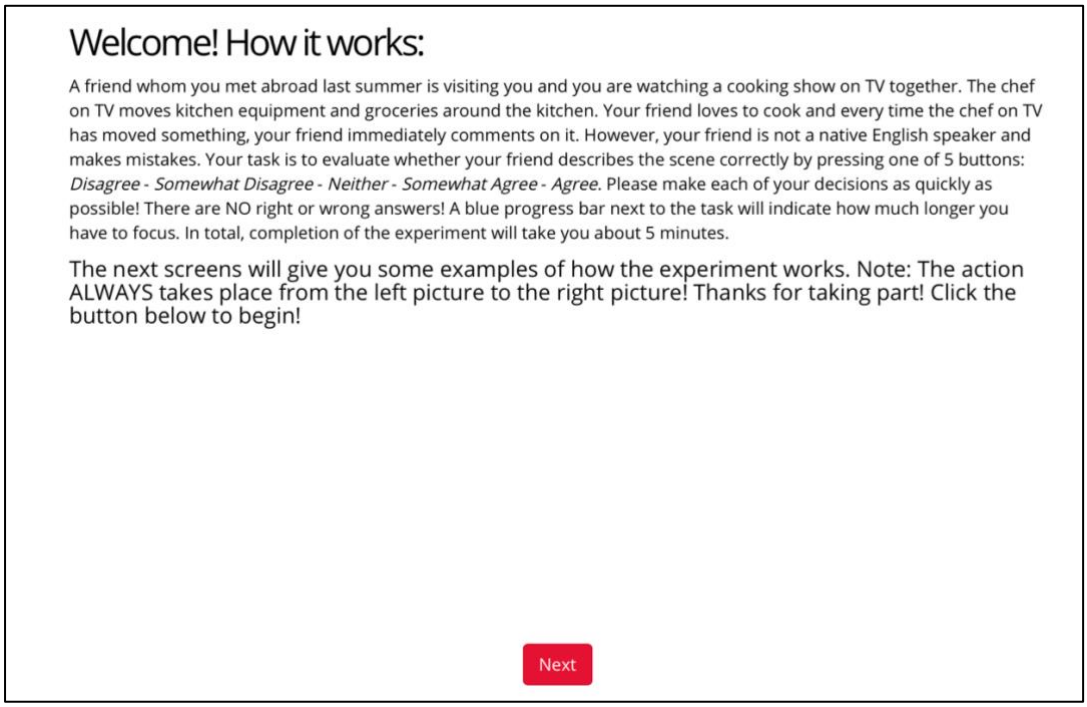
	PLEASE TICK THE FOLLOWING BOXES TO INDICATE YOUR CONSENT:	Tick Box
1.	I hereby confirm that I am 16 years or older.	☐
2.	I confirm that I have read and understood the Participant Information Sheet for the above study. I have had an opportunity to consider the information and what will be expected of me. I have also had the opportunity to ask questions which have been answered to my satisfaction and I would like to take part in the psycholinguistic online experiment.	☐
3.	I consent to participate in the study *"Interpretation of scalar terms by native speakers and foreign language learners"*. I understand that no personal information will be collected.	☐
4.	I understand that no personal information will be collected and that I cannot be identified. I understand that my data gathered in this study will be stored anonymously and securely. My data are associated to an ID that is generated randomly at the beginning of the experiment. I understand that it will not be possible to identify me in any publications, and the researchers cannot have access to my personal information.	☐
5.	I understand that my information may be subject to review by responsible individuals from the University of Oxford for monitoring and audit purposes.	☐
6.	I understand that my participation is voluntary and that I am free to withdraw at any time without giving a reason. I understand that if I decide to withdraw, any personal data I have provided up to that point will be deleted unless I agree otherwise.	☐

7.	I understand the potential risks of participating and the support that will be available to me should I become distressed during the course of the research.	☐
8.	I understand that my participation will not benefit me directly, but will advance the scientific field and may have an impact on educational practices.	☐
9.	I understand that the data will not be made available to any commercial organisations but is solely the responsibility of the researchers undertaking this study.	☐
10.	I hereby confirm that: 6. I am a native speaker of English **or** I am a native speaker of German learning English as a foreign language 7. I am a student at either a German grammar school **or** at a British senior school 8. I declare that I don't have any neurological and learning impairments 9. I have normal or corrected-to-normal vision	☐
11.	I am aware of who I should contact if I wish to lodge a complaint.	☐
12.	I understand that data for this project will be used to advance the scientific field and therefore can be published on peer-reviewed journals. I understand that my anonymised data will be stored indefinitely.	☐

Appendix 5 – Experiment Instructions for Participants

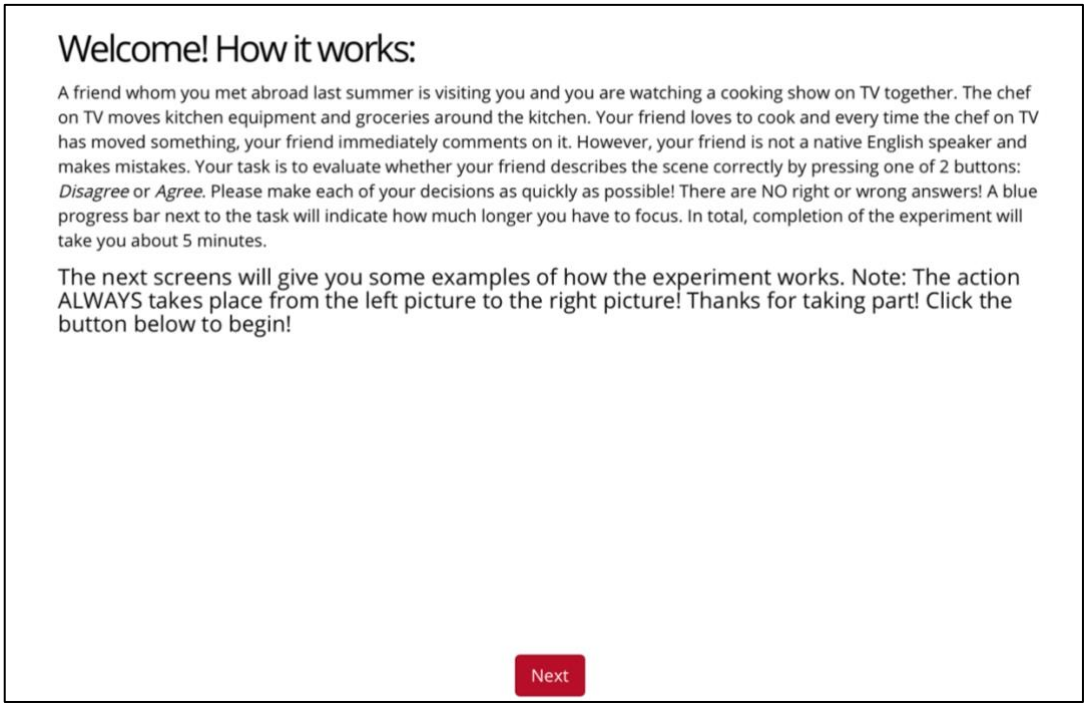
Appendix 5a – Quinary Measure

Fig. A15 Experiment Instructions (quinary measure)



Appendix 5b – Binary Measure

Fig. A16 Experiment Instructions (binary measure)



Appendix 6 – CUREC Approval

Tuesday, March 23, 2021 at 10:21:16 Central European Standard Time

Betreff: CUREC ED-CIA-21-164 - Approval
Datum: Freitag, 19. März 2021 um 12:12:46 Mitteleuropäische Normalzeit
Von: Laura Molway
An: Johannes Schulz, Elizabeth Wonnacott
CC: Student CUREC
Anlagen: image003.png

Dear Johannes,

Title: Interpretation of scalar terms by native speakers and foreign language learners
Ref: ED-CIA-21-164

The above application has been considered on behalf of the Departmental Research Ethics Committee (DREC) in accordance with the procedures laid down by the University for ethical approval of all research involving human participants.

I am pleased to inform you that, on the basis of the information provided to DREC, the proposed research has been judged as meeting appropriate ethical standards, and accordingly, approval has been granted.

Please continue to follow all current guidance issued by CUREC during the pandemic, notably COVID-19: CUREC guidance on research involving human participants, https://researchsupport.admin.ox.ac.uk/governance/ethics/coronavirus. The best practice guidance for internet-based research may also prove useful: https://researchsupport.admin.ox.ac.uk/governance/ethics/resources/bpg

Should there be any subsequent changes to the project which raise ethical issues not covered in the original application you should submit details to student.curec@education.ox.ac.uk for consideration.

Good luck with your research study.

With best wishes,
Laura Molway
Member of DREC

Laura Molway (she/her)
Departmental Lecturer in Modern Languages Education
St Antony's College
@OxfordDeptofEd

Recent work:
Measuring effective teaching: Student perceptions of their modern languages lessons in England
What do languages teachers in England say they want to develop?

Appendix 7 – Normal distribution, visual check, quinary data

Fig. A17 Distribution of mean answer scores of English L1 participants in the underinformative condition in the quinary measure. Each dot represents one participant.

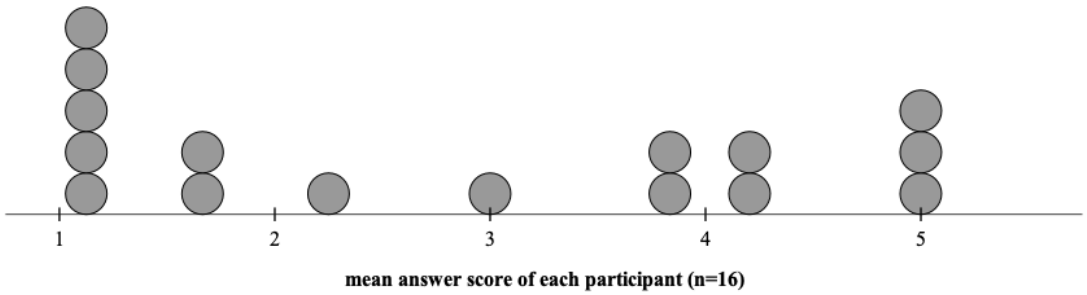
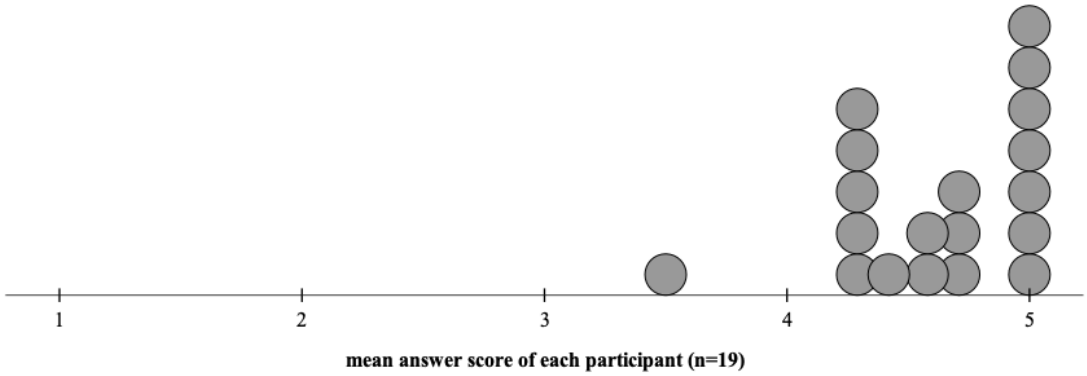


Fig. A18 Distribution of mean answer scores of German EFL participants in the optimally false condition in the quinary measure. Each dot represents one participant.



Appendix 8 – Normal distribution, visual check, binary data

Fig. A19 Distribution of mean answer scores of German EFL participants in the underinformative condition in the binary measure. Each dot represents one participant.

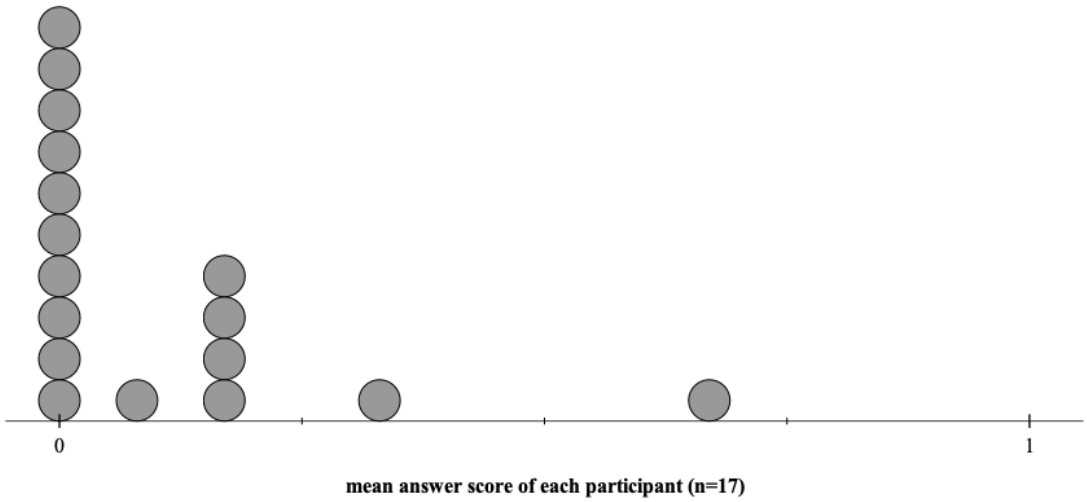


Fig. A20 Distribution of mean answer scores of English L1 participants in the underinformative condition in the binary measure. Each dot represents one participant.