Glossa a journal of general linguistics

Fekete, István, et al. 2018. Exhaustivity in single bare *wh*-questions: A differential-analysis of exhaustivity. *Glossa: a journal of general linguistics* 3(1): 96. 1–32, DOI: https://doi.org/10.5334/gjgl.549

RESEARCH

Exhaustivity in single bare *wh*-questions: A differential-analysis of exhaustivity

István Fekete[1], Petra Schulz[2] and Esther Ruigendijk[1,3]

[1] University of Oldenburg, School of Linguistics and Cultural Studies, Institute of Dutch Studies, 26111 Oldenburg, DE

[2] University of Frankfurt, Institute for Psycholinguistics and Didactics of German, Norbert-Wollheim-Platz 1, 60323 Frankfurt am Main, DE

[3] University of Oldenburg, Cluster of Excellence "Hearing4all", 26122 Oldenburg, DE

Corresponding author: István Fekete (istvan.fekete1@uni-oldenburg.de)

Despite a large body of research, the linguistic nature of exhaustivity in single *wh*-questions is unresolved. Moreover, little empirical evidence exists as to which related structures pattern with bare *wh*-questions regarding exhaustivity. This paper explores the felicity of various exhaustivity violations in unembedded single bare *wh*-questions in German and compares them to related structures. In two novel felicity judgment experiments, a total of 441 participants rated exhaustive as well as non-exhaustive plural and non-exhaustive singleton answers to *wh*-questions or statements in a questionnaire. Answers were based on picture stimuli depicting individuals performing various actions. The felicity of non-exhaustive answers was compared across four main test conditions: bare *wh*-questions (*wer* 'who'), *wh*-questions with a lexical exhaustivity marker (*wer alles* 'who all'), plural definite descriptions contained in a restrictive relative clause (e.g., "the people who are fishing in the garden"), and the scalar quantifier "some" (e.g., "some people who are fishing in the garden").

We employ a novel methodological approach to improve the interpretability of statistical differences between experimental conditions by using the statistical measure of Minimal Important Difference (MID). Our results from estimated MIDs reveal that adults' felicity judgments of non-exhaustive plural answers to bare *wh*-questions pattern with those to *wer alles*-questions and to plural definite descriptions: exhaustivity violations in the bare *wh*, the *wer alles* and the plural definite conditions were rated as less felicitous than exhaustivity violations in the *some*-condition.

Keywords: exhaustivity; semantics; *wh*-question; German; felicity judgment; Minimal Important Difference (MID)

Introduction

Informally speaking, an exhaustive answer is a true and complete answer to a request or a question. A true, exhaustive answer to the request "Name all the students who failed the test", for example, would consist of a complete list of all those students who in a specific situation failed the test. Put differently, an exhaustive answer specifies the maximal set of individuals satisfying the predicate in question, which in our case denotes the property of having failed the test. This answer is maximally informative in that it tells the addressee to which set of possible worlds her actual world belongs (see Zimmermann 2007b).

A true, non-exhaustive answer, in contrast, is not maximally informative. It would consist of a list of just some students who in this situation failed the test. Exhaustivity can be triggered by various structures, *i.a.* clefts, *only*, and *wh*-questions. Focusing on German, in the present study we examine subject *wh*-questions (*wer* 'who'), subject *wh*-questions with the overt exhaustivity marker *alles* (*wer alles* 'who all'), plural definite descriptions contained in a restrictive relative clause, as in "Name the people who", henceforth also referred to as definite plural NPs, and *einige* ('some') in requests.[1]

The aim of this paper is to advance the discussion of the nature of the exhaustivity effect in single bare *wh*-questions. The main question is whether exhaustivity forms part of the inherent meaning of *wh*-questions (e.g., Nelken & Shan 2004) or whether exhaustivity effects with *wh*-questions arise from independent reasons such as application of general pragmatic strengthening processes, which may be triggered by the context, by world knowledge, etc. (e.g., Zimmermann 2010). The former explanation we refer to as the semantic analysis of *wh*-questions and the latter explanation we refer to as the pragmatic analysis of *wh*-questions.

We explore the nature of exhaustivity in single bare subject *wh*-questions by comparing participants' judgments of non-exhaustive and exhaustive answers to *wer*-questions to *wer alles* and plural definite descriptions, which do not allow for non-exhaustive answers, as well as to *einige*, which can prompt a 'some but not all' response.

The picture stimuli we used always showed more than one individual satisfying the predicate in question. Given these more-than-one-individual scenarios, participants had to judge the felicity of the responses that were exhaustive, singleton or non-exhaustive. An exhaustive answer lists all and only those individuals satisfying the predicate in question; a singleton answer contains one individual satisfying the predicate in question, and a non-exhaustive answer contains more than one but not all individuals satisfying the predicate in question.

To our knowledge, this study is the first to assess the felicity of these three answer types in unembedded *wh*-questions, using the Question-about-a-picture task developed by Schulz and colleagues (see Roeper et al. 2007; Schulz & Roeper 2011; Schulz 2015). Moreover, extending the standard ways of statistical analyses, in this study we employ the notion of Minimal Important Difference (henceforth MID), first used by Jaeschke et al. (1989). MID is derived between experimental conditions (i.e. cut-offs) and then compared to the extent of observed differences between the experimental conditions in order to minimize the risk of over-interpreting small, but significant results. Only if an observed difference exceeds the cut-off, is the difference of theoretical importance (Jaeschke et al. 1989). This way erroneous interpretations can be avoided that would attribute linguistic meaningfulness to statistically significant though trivial differences.

The paper is structured as follows: In Section 1 the theoretical background for the current study is provided as well as previous psycholinguistic and neurophysiological research into exhaustivity and relevant methodological considerations. Section 2 and Section 3 contain the two experiments. Section 4 describes the method of estimates of Minimal Important Difference (Jaeschke et al. 1989) for the critical contrasts in our two experiments and draws inferences about the theoretical relevance of differences. The two experiments are discussed in Section 5, focusing on the felicity of incomplete answers to *wer*- and *wer-alles*-questions in comparison to other structures.

1. Background

1.1. Wh-questions – semantic and pragmatic accounts

There is a large body of research on the meaning of single bare *wh*-questions. In this study we focus on who-questions in German (*wer* 'who'), where plurality is unmarked,[2] just like in the English equivalent *who*. Under the so-called "mention-all" reading (Groenendijk &

[1] In this paper we restrict ourselves to unembedded *wh*-questions. For recent accounts on the semantics of embedded *wh*-questions see, for example, Beck & Rullmann (1999), Klinedinst & Rothschild (2011), Romero (2015), and Uegaki (2015).

[2] Note that although the pronoun *wer* is singular, it can be used with a plural noun in a "Gleichsetzungssatz" ("identificational sentence"). In these cases, the noun determines the form of the verb, e.g., "Wer sind die Opfer?" ('Who are the victims?').

Stockhof 1984), the answer to the *wh*-question is exhaustive, i.e. true and complete. The answer has to list all individuals satisfying the predicate in a given world or situation. This is also referred to as a *universal* or *maximal* interpretation (Caponigro et al. 2012). For example, if the judge in a legal trial asks me "Who came by car?", and I know that three individuals were in a car and nobody else, I have to name all three individuals and nobody else, unless I want to tell a lie or choose not to cooperate. Giving the name of only one person, or giving the name of two out of the three individuals, are both non-exhaustive answers. As they are incomplete, albeit true, answers, in this context they would violate the exhaustivity requirement imposed by the legal context.

Besides the "mention-all" reading, single bare *wh*-questions have a "mention-some" reading.[3] If I am hosting a party and one of the guests asks "Who came by car?", because she is looking for a ride home, as the host I will most likely respond by mentioning one or two but not all guests who came by car, i.e. I will answer the question non-exhaustively.

There is consensus that these two readings exist. Different accounts, however, have been proposed as to the linguistic nature of the exhaustivity information triggered by unembedded bare *wh*-questions. Semantic approaches (*i.a.* Karttunen 1977; Groenendijk & Stockhof 1984; Nishigauchi 1999; Nelken & Shan 2004; Schulz & Roeper 2011) and pragmatic approaches (*i.a.* Van Rooij 2003; Schulz & Van Rooij 2006; Zimmermann 2007a) have postulated different mechanisms to account for the nature of exhaustivity information in *wh*-questions.

Semantic analyses propose that exhaustivity/non-exhaustivity, or in other terms universal/existential interpretations (Reich 1997), are a matter of ambiguity between the exhaustive and the non-exhaustive interpretations (e.g., Beck & Rullmann 1999). In some approaches the exhaustive response ("mention-all" reading) is argued to be the default (e.g., Schulz & Roeper 2011). This is in line with Nishigauchi's (1999) assumption that English *wh*-expressions have a covert universal quantifier *every*. On the semantic view, the exhaustive and the non-exhaustive interpretations are inherent grammatical properties of question meaning. The exhaustive meaning involves universal quantification over the individuals in a contextually given domain and the non-exhaustive meaning involves existential quantification over an individual (Nelken & Shan 2004; Roeper et al. 2007). Depending on the situation, one of the two semantic representations is chosen. We refer to this analysis of exhaustivity in *wh*-questions as semantic.

In contrast to semantic accounts, in pragmatic accounts exhaustivity is not part of the semantic meaning of a single *wh*-question (e.g., Van Rooij 2003; Schulz & Van Rooij 2006; Zimmermann 2007a). Taking Van Rooij's account (2003), exhaustivity inference is a pragmatic mechanism triggering the smallest set of candidates that gives optimal relevance depending on the context. The question "Who came by car?", for example, is semantically underspecified and the inference denotes the optimal answer (i.e. a certain number of people) satisfying the predicate in question by considering the context of the person asking for this information. Van Rooij (2003) derives relevance as a decision problem by taking into account the goal of the information exchange. He incorporates the decision problem of the questioner as a contextual parameter that resolves a question. The answer should be useful to resolve a question. Crucially, resolvedness is tied to the goals of the questioner. Under this account, the "mention-some" reading can be derived if all true possible answers are equally useful for the questioner.

Previous empirical research has shown that exhaustivity violations of semantic nature, such as with the focus particle *only*, where exhaustivity is part of its truth-functional meaning,

[3] Recent analyses allow for "mention-some" reading in the presence of possibility modals only (Xiang & Cremers 2017).

can be discerned from those of non-truth-functional nature based on the perceived degree of violation (Drenhaus et al. 2011, see also Section 1.2). Exhaustivity violations of non-truth-functional nature are perceived as less severe than exhaustivity violations of semantic (truth-functional) nature. Severity can be operationalized as the response strength of the rejection on the Likert-scale.

Our two large-scale offline questionnaire studies explore the linguistic nature of exhaustivity in bare single *wh*-questions by comparing these to related structures that only license exhaustive answers (*wer alles* 'who all', plural definite descriptions) and to a structure that can have both an exhaustive and a non-exhaustive answer (*einige* 'some').

1.2. Related and contrasting structures involving exhaustivity

This section outlines structures that have been argued to be similar to or to differ from bare *wh*-questions regarding exhaustivity: *wer alles* ('who all'), *wer so alles* ('who all'), the plural definite description triggered by *die*, and the scalar quantifier *einige* ('some').

Most closely related to bare *wh*-questions are overtly marked *wer alles*-questions ('who all'). It has been suggested that *alles* ('all') has its own semantics and directly adds to propositional content of the question (Zimmermann 2007a; b), resulting in a weakly exhaustive interpretation, i.e., referring to the conjunction of all propositions that are true (Beck & Rullmann 1999: 288). An exhaustive list answer to the question "Who all is fishing?" is a case of weak exhaustivity, because we say who is fishing but not who is not fishing. Strong exhaustivity, which would entail that we also know who in a given discourse is not satisfying the predicate in question, has been argued to be triggered in *wh*-questions embedded under *know* (Groenendijk & Stokhof 1984). For example, the propositions "John knows who was at the party" and "Mary was not at the party" license the inference that "John knows that Mary was not at the party".

Besides *wer alles,* plural definite descriptions contained in a restrictive relative clause served as our second semantic condition for the following reasons. In the statement "The people she invited were students", for example, the plural definite description *the people* denotes all the individuals out of the set of relevant individuals who satisfy the predicate "being invited". Link (1983) proposes a semantic analysis of plural definite descriptions that uses a mereological fusion operator: $[|\text{the NP}_{\text{pl}}|] = \sigma x\ [\text{NP}(x)]$, meaning 'the sum of all x that are NP'. A non-exhaustive interpretation would hence constitute a semantic violation of the exhaustivity requirement of the determiner phrase (DP) *the people* (Sharvy 1980; Link 1983; Schwarz 2013). Note, however, that other approaches suggest that plural definite descriptions are underspecified and that maximality (i.e. exhaustivity) is triggered pragmatically (e.g., Sauerland et al. 2005; Malamud 2012). In this vein, plural definite descriptions can license a non-maximal interpretation, for example, in the sentence "The windows are open", which may refer only to a subset of windows (Malamud 2012: 11). Because of the attested uncertainty about the semantic or pragmatic nature of maximality in plural definites and their possible non-exhaustive interpretation, we used plural definite NPs modified by restrictive relative clauses, as in "Name the people who are fishing in the garden". This sentence represents a clear condition of semantic exhaustivity, because the predicate of the relative clause restricts the set of individuals to the maximum by modifying the noun (Partee 1975). The semantic mechanism is such that the restrictive relative clause, which denotes the set of contextually relevant individuals who are satisfying the predicate, combines with the set of individuals that the plural definite description designates by predicate modification (Montague 1973; Heim & Kratzer 1998).

The structure with *einige* ('some') was chosen because, just like *wer*, it has two interpretations. *Wer* has an exhaustive and a non-exhaustive interpretation, and *some* has the readings: 'some but not all' and 'some and in fact all'. 'Some but not all' is called the scalar implicature of "some". Since Grice (1989), this pragmatically enriched inference 'not all' has been analyzed to result from reasoning about the intention of the speaker's utterance: the speaker chose "some" instead of "all" deliberately to indicate that the latter does not hold. Imagine someone asks "Have the students passed the exam?", and the answer is "Some (of them) have." In this case the scalar implicature is clearly activated, and "some" does not have the reading 'some and in fact all the students'. In contrast, the sentence "Eating some of these pills will kill you" is clearly compatible with the reading 'eating all of the pills will kill you'. In this case, the logical-semantic meaning of "some" ('at least one and in fact all') is activated.[4]

Generally speaking, the choice between the implicature and the logico-semantic reading of "some" primarily depends on observing the maxims of Cooperation and Quantity that guide conversational efficacy (Grice 1989). However, other factors can also play a role in the generation of implicatures, *i.a.*, enriched experimental instructions (Papafragou & Musolino 2003; Guasti et al. 2005), the task context (Politzer-Ahles & Fiorentino 2013; Degen & Goodman 2014; Degen & Tanenhaus 2015), the grammatical structure (Hartshorne et al. 2015), the methodological design (Degen & Tanenhaus 2011; 2015), individual background factors, such as unique interpretive preferences, cognitive measures (Dieussaert et al. 2011; Marty & Chemla 2013; Antoniou et al. 2016), bilingualism (Syrett et al. 2017a; b; Dupuy et al. 2018), L2 acquisition (Miller et al. 2015; Snape & Hosoi 2018), neurological conditions (Kasher et al. 1999; Pastor-Cerezuela et al. 2018), socio-economic status (Wilson 2017), socio-pragmatic abilities (Nieuwland et al. 2010; Zhao et al. 2015; Barbet & Thierry 2016) or theory of mind capacity to infer the intention of the speaker (Cummings 2015). In light of these potential factors influencing the derivation of implicatures, we selected our sample of participants as homogenous as possible, recruiting only native speakers of German who are university students. We elaborate on the task-related factors in Section 5.

Importantly, given our experimental scenarios (see. Section 2.1.2 and Section 3.1.2), for *wh*-questions the non-exhaustive answer is considered a violation, because as a partial answer it is under-informative, while for *einige* the exhaustive response is considered a violation, because it is over-informative. Note that in this paper, we focus on the weakly exhaustive meaning of *wh*-questions.

Assuming that exhaustivity is specified semantically as proposed *i.a.* by Schulz & Roeper (2011), in the specific experimental set-up in our study a bare *wh*-question such as "Who is fishing in the pond?" would hence be tantamount to asking "Who all is fishing in the pond?" or "Name everybody who is fishing in the pond". Therefore, we assume that infelicitous, non-exhaustive answers to bare *wh*-questions would pattern with non-exhaustive answers in the *wer alles*-condition and in the plural definite description condition.

In the case of *einige*, one of the meanings represents the semantic one ('some and in fact all'), and the other one the pragmatic one ('some but not all') (Horn 1972; Grice 1975; 1989). This is in contrast to *wh*-questions, where, as mentioned above, both the non-exhaustive and the exhaustive meaning are argued to be represented semantically (Nelken & Shan 2004). We expected both non-exhaustive and exhaustive responses to *einige* to be perceived as more felicitous than non-exhaustive responses in the critical conditions of *wer*, *wer alles,* and the plural definite description.

[4] Note that in contrast to pragmatic analyses of scalar implicatures, Chierchia et al. (2012) suggest a semantic operation, criticizing (neo)-Gricean pragmatic accounts. More defenses in favor of Gricean accounts of scalar implicatures are found in Russell (2006).

1.3. Previous psycholinguistic and neurophysiological research into exhaustivity

Exhaustivity has been studied in adults using acceptability judgements (Drenhaus et al. 2011), picture-sentence-verification experiments (e.g., Gerőcs et al. 2014) as well as neurophysiological and questionnaire experiments (Drenhaus et al. 2011). Drenhaus et al. (2011) investigated the nature of exhaustivity in German *it*-clefts comparing them to *only*-foci using an offline acceptability judgement task with a 1–6 Likert-scale complemented by an ERP study (Event Related Potentials).

Drenhaus et al. (2011) explored whether the nature and source of exhaustiveness effects with *it*-clefts are semantic (i.e. truth-functional) in nature or whether they are non-truth-functional (i.e. as a presupposition or a generalized conversational implicature). To this end, they first compared exhaustivity violations in *it*-clefts and *only*-foci in a questionnaire study. Examples for their test sentences are given in the examples (1) and (2) below (Drenhaus et al. 2011: 7):

(1) a. [*it*-cleft, [+exh]]
Es ist Maria, die das Klavier spielen kann und außerdem noch die
it is Maria that the piano play can and besides also the
Geige, sagte…
violin said…
'It is Mary that plays the piano and, besides, the violin, said…'

b. [*it*-cleft, [-exh]]
Es ist Maria, die das Klavier spielen kann und außerdem noch Luise
it is Maria that the piano play can and besides also Luise
und Jana, sagte…
and Jana, said…
'It is Mary that plays the piano and, besides, Luise and Jana, said…'

(2) a. [*only*-focus, [+exh]]
Nur Maria kann das Klavier spielen und außerdem noch die
only Maria can the piano play and besides also the
Geige, sagte…
violin said
'Only Mary can play the piano and, besides, the violin, said…'

b. [*only*-focus, [-exh]]
Nur Maria kann das Klavier spielen und außerdem noch Luise
only Maria can the piano Play and besides also Luise
und Jana, sagte
and Jana said
'Only Mary can play the piano and, besides, Luise and Jana, said'

The participants were asked to rate the acceptability of sentences such as (1) and (2). Exhaustivity violation of *it*-clefts (Mean rating = 2.8) differed significantly from that of *only*-focus (Mean rating = 3.7). *Only*-focus triggers exhaustivity semantically as "part of the asserted truth-functional content of the utterance" (Drenhaus et al. 2011: 2), whereas for *it*-clefts, the question arises whether the *exhaustiveness effect* is derived truth-functionally via the propositional content as well. The authors found a mean difference of 0.9 on a scale of 1–6 between mean acceptability ratings of exhaustivity violations of *only*-foci and *it*-clefts. They concluded that given this difference, exhaustivity in *it*-clefts – unlike in *only*-foci – is not a truth-functional phenomenon. Importantly, the comparison of the extent of violations across the two structures allowed them to draw inferences about the linguistic nature of *it*-clefts. Note that there was no non-truth-functional, i.e. pragmatic, control condition in the experiment to compare *it*-clefts to clearly non-truth-functional violations.

Drawing on this difference between truth-functional and non-truth-functional violations, Drenhaus and colleagues then developed an online experiment, using ERPs, to confirm the qualitative nature of the difference revealed in their questionnaire study. ERPs are the measured brain responses to specific sensory stimuli (for a review on ERPs and language-processing, see Kaan 2007). We can distinguish different ERP waveforms. N400, one of the best-studied language-related ERP waveforms, is a negative-going wave 300–600 ms post-stimulus. It was first described by Kutas & Hillyard (1980) as a marker of semantic incongruity, as illustrated by the last word in the sentence "He took a sip from the transmitter" (example in Kutas & Hillyard 1980: 203). N400 has also been found for contextually-induced failed expectations and it is linked to the strength of lexical associations (for a review of N400, see Lau et al. 2008).

Another important effect related to language processing is the so-called P600. This is a relatively late (between around 500–1000 ms after the stimulus) positive deflection in the EEG signal, which was originally analyzed as a syntactic component (Hagoort et al. 1999; Friederici & Weissenborn 2007; Gouvea et al. 2010). It emerges in native speakers as a result of syntactic violation, syntactic complexity, syntactic repair/reanalysis or an unexpected syntactic structure (Osterhout & Holcomb 1992). Both N400 and P600 have been shown to be elicited by a broader range of phenomena (see e.g., Van Herten et al. 2005; Politzer-Ahles et al. 2013; Brouwer et al. 2017). P600, for example, has also been found in logical-semantic violations of licensing of negative polarity items in German (Drenhaus et al. 2005; 2006). Furthermore, even though truth-value violations are semantic violations, N400 has not been found in truth-value violations with semantically related words such as "A robin is not a bird" (Fischler et al. 1983). Based on these two findings, Drenhaus et al. (2011) expected a P600 to exhaustivity violations if exhaustivity is "a conventionalized part of the meaning of these structures" (Drenhaus et al. 2011: 6). In line with these predictions, they found an N400 effect for non-exhaustive *it*-clefts but not for exhaustive *it*-clefts, as well as a P600 effect for non-exhaustive *only*-sentences but not for exhaustive *only*-sentences. These findings support the conclusion based on the offline measure. For our study, we used Drenhaus et al.'s (2011) line of reasoning and compare the extent of perceived infelicity in several conditions to see with which other structures *wh*-questions pattern.

1.4. Methodological considerations

In Experiments 1 and 2, we employed two types of exhaustivity violations, modelled as answer types, across different structures: non-exhaustive (i.e. naming several but not all) and singleton exhaustivity violations. The labels refer to the number of individuals listed in the answer: in the case of the singleton only one of several individuals is mentioned, in the case of the non-exhaustive more than one but not all individuals satisfying the predicate are mentioned. The singleton condition was added to provide participants with varied response types. We are aware that a singleton answer in case of the plural definite descriptions constitutes a double violation: it is an exhaustivity violation and a violation of number agreement. As we were primarily interested in the non-exhaustive condition, which does not violate number agreement, we carried out planned comparisons between structures for the singleton and for the non-exhaustive level of violation separately.

We created pictures of garden scenes and sentence stimuli, which were embedded in an explicit task context (see Section 2). The NP *im Garten* ('in the garden') was added to all verbal test stimuli to make explicit that the *wh*-question is asked about the specific picture only. Furthermore, we explained in the instruction that both the person asking the question and the one giving the response are looking at the same picture scene, i.e. share common knowledge.

A further point of concern is that the structures in our conditions inevitably differ regarding the overtness of the exhaustivity marking, as for instance in *wer alles* vs. *wer*. The lexical exhaustivity marker *alles* may highlight exhaustivity more than the covert exhaustivity operator in the unmarked condition *wer*, and this may result in statistically significant differences given our large sample sizes. Since statistical significance is influenced by the number of observations, significant comparisons could be observed even though differences are small. Generally speaking, large-scale studies may lead to statistically significant results without indicating linguistically meaningful differences (see Lin et al. 2013; Angst et al. 2017).

To address this concern, we operationalized linguistic meaningfulness using the MID. Observed means in the same neighborhood, i.e. pairwise differences within the bounds of MID, are regarded as linguistically not meaningful differences, while observed mean differences exceeding the MID-threshold are considered linguistically meaningful differences. In Section 4 we describe the statistical analysis methods using MIDs and discuss the extent of observed differences.

2. Experiment 1

Experiments 1 and 2 examine the extent of an exhaustivity violation of bare *wh*-questions in relation to other structures. Primary statistical results of the effects of the two experiments are presented with the experiments; secondary statistical results of MID estimates from the two experiments are compared in Section 4.1 and Section 4.2. Section 5 discusses the results of the two experiments.

Experiment 1 explored the felicity of non-exhaustive answers in the following Sentence Type conditions: *wer, wer alles*, *wer so alles*, the plural definite description triggered by the relative-pronoun *die*, and *einige*. We are mainly interested whether non-exhaustive answers to the bare *wh*-question *wer* ('who') pattern with those to *wer alles*, the plural definite description (semantic violation) or with those to *einige*. The structure *wer so alles* was included in Experiment 1 for exploratory reasons to examine the meaning of this structure in terms of exhaustivity in relation to the meaning of the *wh*-questions and *wer alles*-questions.

We adopted the Question-about-a-picture exhaustivity task, which was developed by Schulz & Roeper (2011) for language acquisition (first used by Roeper & de Villiers 1991 and de Villiers & Roeper 1993 with *wh*-questions). Exhaustivity violations in our experiments were operationalized as degrees of incompleteness of the answer given the scenario depicted in the picture. Following Schulz & Roeper (2011) and Schulz (2015), our picture stimuli depicted a hypothetical family whose members were engaged in various actions in the pictures. However, we made an important modification: rather than eliciting answers to the *wh*-questions, the participants had to judge the felicity of answers on a Likert-scale. Using a Likert-scale allowed us to evaluate fine-grained differences regarding the degree of felicity.

2.1. Method

2.1.1. Participants

German native speakers were recruited from the University of Oldenburg, other universities and via the internet. For both Experiments 1 and 2, informed consents were obtained from the participants before their participation. Four vouchers were drawn randomly and distributed among those who participated. The participants were assigned randomly to one of the counter-balance lists of the questionnaire, so that a close to equal number of participants were tested with every list. Incomplete questionnaires or those filled in by non-native speakers were excluded from the experiment (around 10% of the all questionnaires were discarded as a result of these criteria). Altogether 134 complete questionnaires were analyzed (20 males and 114 females, aged from 17 to 57, $Mean = 24.16$, $SD = 6.56$).

2.1.2. Material

The questionnaire comprised 100 questions in German: 44 critical items, 50 filler items and 6 practice items. The critical items were created with verbs that do not require internal arguments, for example, *baden* ('to bathe', see Appendix B for a complete list). The filler sentences were created to elicit negative judgments, i.e., by naming a wrong person performing the action. The verbs in both critical and filler trials were used in questions such as "Wer steht im Garten?" ('Who is standing in the garden?'), and for two conditions (*einige* and the plural definite description) in affirmative sentences that functioned as requests.

All test structures were followed by singleton, non-exhaustive or exhaustive answers (see Table 1). Question stimulus and answer were presented below each other in dialogue-style (see Figure 1). For reasons of comparability of our results with the study by Drenhaus et al. (2011), we adopted the same 6-point rating scale, however, we used a reversed scale (1: very bad, 6: very good) to distinguish the scale from that of the German school grading system. A further reason for employing the 6-point scale was to not offer a middle value.[5]

The answers to the questions that the participants had to judge were labeled by their level of exhaustivity: singleton, non-exhaustive, and exhaustive. In the case of plural definite descriptions and *wer alles*, the singleton and the non-exhaustive answers we considered infelicitous. An incomplete answer to a *wh*-question, or an over- or under-informative answer to *some* are still true – albeit not necessarily felicitous – answers.

The last DP in the answer was always the same within an item-block. To illustrate, the DP *Der Junge* ('the boy') as a singleton answer is also the last DP in the plural non-exhaustive answer ("Die Mutter und der Junge", 'the mother and the boy') as well as in the exhaustive answer ("Die Mutter, der Vater und der Junge", 'the mother, the father and the boy'). The questionnaire was checked with native speakers of German, who judged both the critical and the filler dialogues on naturalness and grammaticality.[6] Our setup gave rise to 11 experimental conditions, as illustrated in Table 1. We were primarily interested in plural violations of exhaustivity, therefore the singleton conditions were included in a subset of the conditions to restrict the number of conditions in the design.

Table 1 illustrates all conditions using the fishing-item as an example. The garden-scene in the example trial comprised the grandfather, the grandmother and boy fishing in the picture, father taking a photo, girl ironing, and mother digging.

Plural definite descriptions were contained in relative clauses with the nouns *Personen* ('persons'), *Menschen* ('people'), and *Leute* ('people') to create variability. Each participant read 100 items, which were presented in pseudo-random order to control for order effects. Each participant read one particular item only per condition. This set-up gave rise to 11 Latin-square pseudo-randomized counter-balance lists, which were entered as a random variable in the statistical analyses. Each participant saw only one list. In addition, each verb-item was presented with every Sentence Type Condition across the experiment. Each

[5] As pointed out by one of the reviewers, even point-scales reflect both gradience and categorization: a participant response in the range of 1–2 can be considered one category, 3–4 a middle category, and 4–6 another one. Responses at the ends of the scale (1 or 6) can generally be considered as expressing a stronger opinion or as expressing stronger certainty regarding the felicity than responses in the middle (3 or 4), which can be argued to reflect less strong opinions or less certainty on felicity.

[6] Certain items were discarded in the pre-selection phase of the experiment, for example, *grillen* ('grilling') because even the non-exhaustive scenario can be understood as exhaustive (people standing around the grill can also be considered as grilling passively).

Table 1: The 11 critical conditions in the questionnaire in Experiment 1.

Sentence Type	Example for Question	Answer Type	Answer to judge
wer-question	Wer angelt im Garten? who fishes in.the garden 'Who is fishing in the garden?'	Exhaustive	Der Großpapa, die Großmutter und der Junge. the grandpa the grandma and the boy 'The grandpa, the grandma and the boy'
		Non-exhaustive	Der Großpapa und der Junge. the grandpa and the boy 'The grandpa and the boy'
wer alles-question (semantic condition)	Wer angelt alles im Garten? who fishes all in.the garden 'Who all is fishing in the garden?'	Exhaustive	Der Großpapa, die Großmutter und der Junge. the grandpa the grandma and the boy 'The grandpa, the grandma and the boy'
		Singleton	Der Junge. the boy 'the boy'
wer so alles-question	Wer angelt so alles im Garten? who fishes so all in.the garden 'Who all is fishing in the garden?'	Exhaustive	Der Großpapa, die Großmutter und der Junge. the grandpa the grandma and the boy 'The grandpa, the grandma and the boy'
		Non-exhaustive	Der Großpapa und der Junge. the grandpa and the boy 'The grandpa and the boy'
		Singleton	Der Junge. the boy 'the boy'
einige (pragmatic condition)	Nenne einige Menschen, die im Garten angeln. name some people the in.the garden fish 'Name some people who are fishing in the garden.'	Exhaustive	Der Großpapa, die Großmutter und der Junge. the grandpa the grandma and the boy 'The grandpa, the grandma and the boy'
		Non-exhaustive	Der Großpapa und der Junge. the grandpa and the boy 'The grandpa and the boy'
die (plural definite descriptions, semantic condition)	Nenne die Menschen, die im Garten angeln. name the people that.PL in.the garden fish 'Name the people who are fishing in the garden.'	Exhaustive	Der Großpapa, die Großmutter und der Junge. the grandpa the grandma and the boy 'The grandpa, the grandma and the boy'
		Non-exhaustive	Der Großpapa und der Junge. the grandpa and the boy 'The grandpa and the boy'

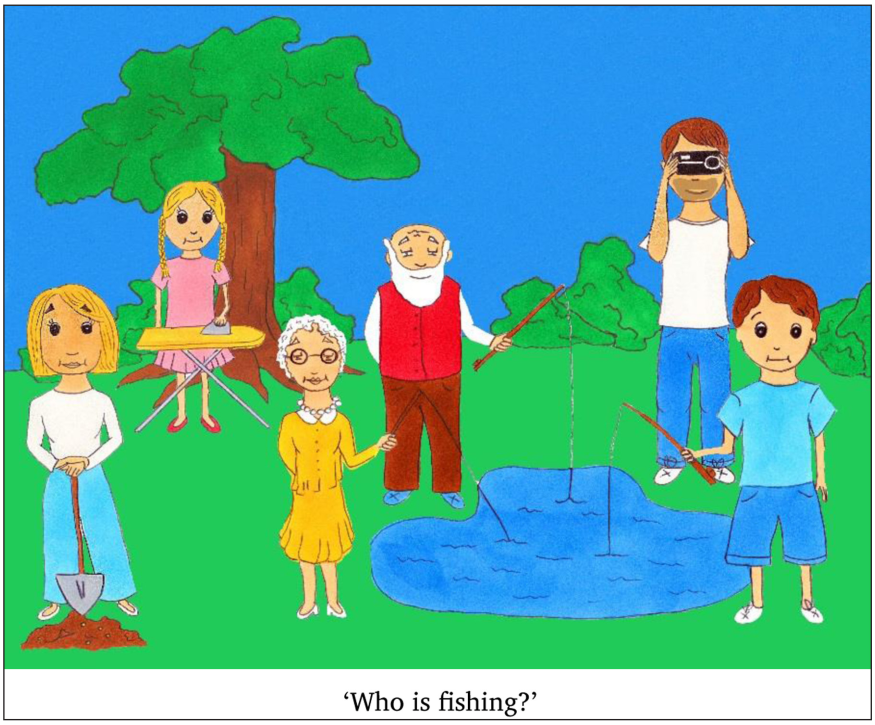

Figure 1: An example of a typical critical picture stimulus in Experiment 1.

participant received four items per condition, which adds up to 44 critical items per participant. This way we made sure that all the items occurred equally frequently in all conditions.

An artist drew the picture stimuli, which illustrated a garden scene with the six family members engaged in different activities. Members of the family were easily distinguishable from one another, for example, the father having a beard, the grandma having glasses, the family members having different clothes on (see Figure 1). The activity referred to by the verb was carried out in the critical trials by three people of the six. In filler trials less or more than three people were carrying out actions. The family members, their clothes, the garden and the scene were kept constant across the picture stimuli to increase coherence across the pictures. In every picture, each of the family members was carrying out one action. When more people were carrying out the same action, those people were displayed as, so they did not form a group. Stereotypical associations such as women knitting or children hopping were avoided. Figure 1 illustrates a typical picture stimulus in the two studies.

2.1.3. Procedure

Participants filled in the questionnaire on the ThesisTools homepage (http://www.thesistools.de). Participants first read a welcome text introducing them to the study. They were unaware of the aim of the study. After answering questions about their gender and age, they read an instruction text. They were told that they would read short dialogues between a hypothetical teacher and a pupil. Each dialogue consisted of a question asked by the teacher about a picture and an answer given by the pupil.[7] We chose this school-scenario

[7] Note that as participants read the dialogues, sentence intonation and inner prosody cannot be controlled for. Fodor (1998; 2002) has argued that in silent reading a default prosodic contour facilitates the resolution of syntactic ambiguity such as relative clause attachment. Zhou et al. (2012) have shown that prosody is used to resolve pragmatic ambiguity in both children and adults. Similarly, in our study prosody may have affected the interpretation (i.e. implicature derivation), especially in the case of "some". Recent work (Tomlinson et al. 2017) has shown that intonation facilitates implicatures in the case of "some".

to provide a setting that enables ratings of responses. Participants were told that both the teacher and pupil saw the same picture.

The participants' task was to judge the answer of the pupil in the dialogue on a scale of 1–6 on felicity given the picture (see the instructions in Appendix A). The specific task was to judge how felicitous the answer of the pupil is. Before the experimental trials, participants saw a picture of the family members and six pictures of each of them alone (see also Schulz 2015). Following this, participants completed five practice trials and then the real experiment started with the critical trials.

2.1.4. Primary Statistical Analyses: Cumulative Link Mixed Models (CLMM)

Standard descriptive analyses were summarized as estimated means, medians, standard deviations (SD), standard errors (SE) and 95% confidence intervals of the means (CI) for the conditions. For the statistical analyses, we used RStudio 1.0.136 (RStudio 2012) built on the R platform (R Development Core Team 2013, version 3.3.2). In a recent Monte Carlo study, Kromrey & Hogarty (1998) suggest that cumulative logit models tend to be more powerful than parametric tests. Therefore, we performed Cumulative Link Mixed Models (CLMM) on our ordinal-scaled data to model both participant- and item-variability (Agresti 2002) using the R package "ordinal" (Christensen 2015).

The raw scores were entered as primary outcome measures (i.e. item ratings per participant and condition) into the statistical analyses. The raw scores were not standardized because CLMMs take inter-participant variation into consideration. Sentence Type was treated as fixed factor in the model with participant, item and list as possible random factors. Inclusion of random- and fixed factors were assessed by comparing the Akaike-Information-Criterion values of fitted models (AIC; Akaike 1974) using Likelihood ratio tests. A decrease of more than 2 values in AIC indicates that the goodness of fit of the model improves significantly (Akaike 1974). We first started out with a model that had a single random factor, and then added additional random factors and random slopes. Likelihood-ratio tests of fixed factors were performed on two minimally different models, while keeping the random-effects structure identical (Barr et al. 2013). P-values were obtained by likelihood ratio tests of the model with the effect in question against the model without the effect in question. The CLMMs were fitted with the Laplace approximation (Pinheiro & Bates 1995) using the "probit" link function and the "equidistant" threshold option. In Experiment 1, the CLMM with Sentence Type contained the conditions *wer*, *wer so alles*, *einige* and plural definite descriptions as levels of Sentence Type and only *non-exhaustive and exhaustive answers* as levels of Answer Type to achieve a balanced design. In Experiment 2, we built a full-fledged CLMM with Sentence Type, Answer Type and their interaction.

Multiple comparisons for ordinal data were computed using the R-package "rcompanion" 1.2.0 (Mangiafico 2015) to compute all the possible pairwise comparisons. To adjust for multiple comparisons, the Benjamini–Hochberg method (Benjamini & Hochberg 1995) was applied under the name of FDR in "rcompanion". This method, which is an extension of the Bonferroni method, controls the false discovery rate (FDR) by attempting to limit the probability of even one false discovery (Mangiafico 2015).

2.2. Results

Table 2 illustrates the estimated means and medians of scores per Sentence Type and Answer Type with standard deviations (SD), standard errors (SE) and 95% confidence intervals of the mean (CI). Estimations are computed from 536 items (4 items per condition per participant * 134 participants).

Median scores in all exhaustive conditions were 6, in the non-exhaustive conditions 4 with one exception (*einige*), and 2 and 3 for the *wer alles* and *wer so alles singleton* conditions, respectively. SD inspection based on Table 2 shows that scores associated to infelicitous responses are more spread than the ratings of felicitous responses. Figure 2 illustrates the distribution of responses visually in the critical Sentence Type conditions as a function of Answer Type.

Table 2: Estimated means and medians of all the critical conditions in Experiment 1.[8]

Sentence Type	Answer Type	Mean	Median	SD	SE	95% CI
wer	Exhaustive	5.903	6	0.506	0.022	0.043
	Non-exhaustive	4.062	4	1.158	0.050	0.098
wer so alles	Exhaustive	5.937	6	0.356	0.015	0.030
	Non-exhaustive	3.858	4	1.153	0.050	0.098
	Singleton	2.784	3	1.193	0.052	0.101
wer alles	Exhaustive	5.897	6	0.556	0.024	0.047
	Singleton	2.502	2	1.073	0.046	0.091
die[9]	Exhaustive	5.909	6	0.505	0.022	0.043
	Non-exhaustive	3.696	4	1.164	0.050	0.099
einige	Exhaustive	5.743	6	0.651	0.028	0.055
	Non-exhaustive	5.112	5	1.122	0.048	0.095

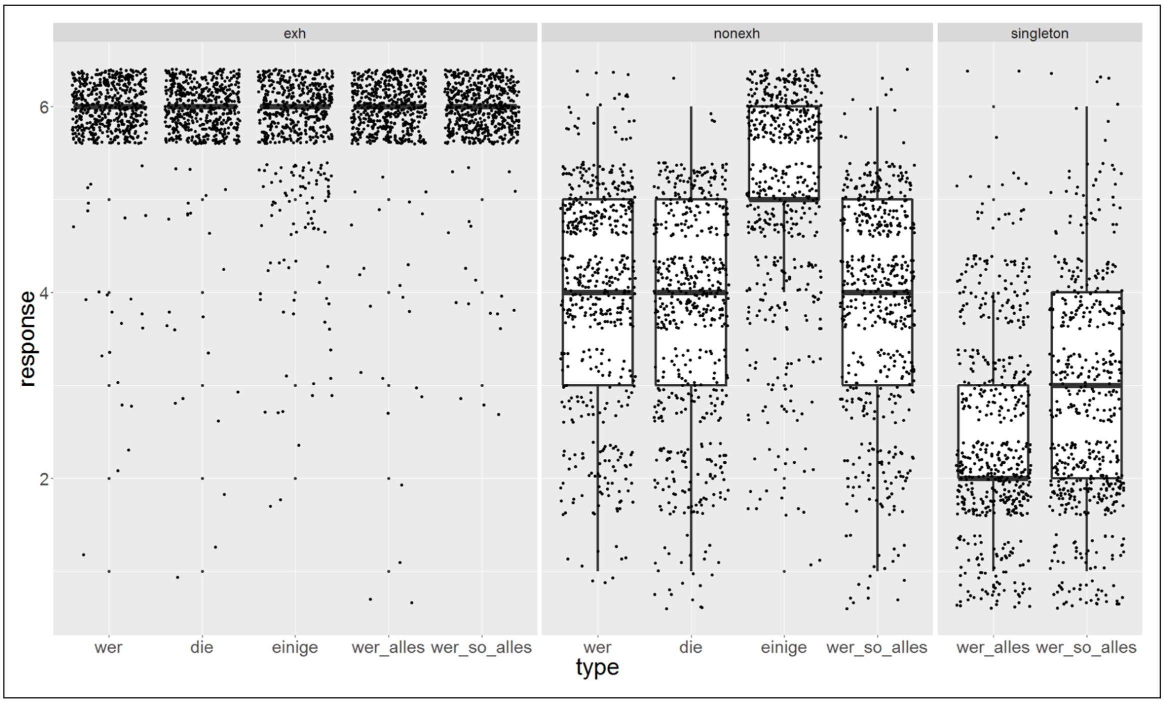

Figure 2: "Jitter" plot with the distribution of responses in the critical conditions in Experiment 1.[10]

[8] *SD* denotes standard deviations, *SE* standard errors of the mean, and *CI* indicates 95% confidence intervals of the mean. *Die* denotes plural definite descriptions triggered by the determiner *die*. Non-exhaustive responses enumerated 2 out of 3 people in Experiment 1.

[9] Plural definite descriptions.

[10] Mean felicity ratings are aggregated by Picture Type (exhaustive, non-exhaustive and singleton) and Sentence Type (*wer*, *wer alles*, *wer so alles*, *einige* and *die* – plural definite descriptions contained in restrictive relative clauses). By visualizing data with jitter-plots, we can unravel patterns in the underlying data. The jitter-plot adds a small amount of random noise to the data to avoid over-plotting, i.e., the overlap of data points.

Figure 2 shows that the rating task leads to similar variance on the conditions, that is, the effect of different interpretations and strategies are constant across the conditions. In other words, the conditions *wer, wer alles, wer so alles* and plural definite descriptions (*die*) are comparable in terms of the distribution of responses along the scale. However, there is a visible cloud of jitter around score 6 in the *wer*-condition. In 107 of 134 cases (80%) the mean within-participant difference between the non-exhaustive *wer*- and non-exhaustive *die*-conditions is smaller than 0.25 (with *wer* being more felicitous), while the group-level mean difference between these conditions (aggregated over participants) is 0.366.

A CLMM with Sentence Type using only the set of *non-exhaustive* responses was computed. List as a random factor did not increase explanatory power of the model as revealed by model comparisons. The model fitted with Sentence Type yielded a better model than the intercept-only model, as measured by the decrease of at least two points in the AIC values ($LR = 4.4339, df = 1, p = 0.035$). This shows that Sentence Type is a significant predictor ($\beta = 0.1294, se = 0.0612, Z = 2.114, p = 0.0345$), that is, the type of the question affected plural non-exhaustive responses, causing a variation over the Sentence Type conditions. Results from pairwise comparisons between Sentence Type conditions will be presented in Section 4.1 in order to be able to directly compare them to those from Experiment 2.

3. Experiment 2

The general aim of Experiment 2 was to improve the design of Experiment 1 by balancing the Picture Type conditions and increasing the number of individuals in the picture stimuli by one. The latter modification was necessary because some respondents in Experiment 1 reported that they found the use of *einige* referring only to 3 people problematic. Similarly, Degen & Tanenhaus (2015) found that *some* is less natural for reference to small sets (1, 2 and 3) compared to intermediate sets (6–8).

Therefore, we included pictures with four people performing an action (see Figure 3). The second modification was that we designed a full-fledged condition set to allow for

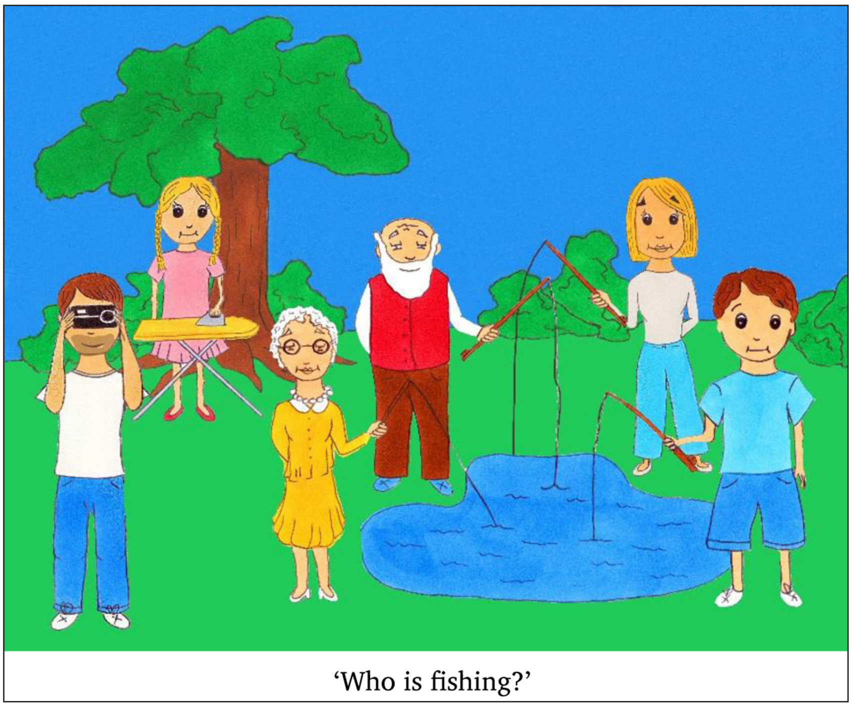

Figure 3: An example of a picture stimulus in Experiment 2. Four individuals are carrying out the critical action (fishing), while the other two persons are engaged in different actions.

Table 3: The 12 critical conditions in the questionnaire in Experiment 2.

Sentence Type	Example for Question	Answer Type	Answer to judge
wer-question	Wer angelt im Garten? who fishes in.the garden 'Who is fishing in the garden?'	Exhaustive	Der Großpapa, die Großmutter, die Mutter und der Junge. the grandpa the grandma and mother and the Boy 'The grandpa, the grandma, the mother and the boy'
		Non-exhaustive	Der Großpapa, die Großmutter und der Junge. the grandpa the grandma and the boy 'The grandpa and the boy'
		Singleton	Der Junge. the boy 'the boy'
wer alles-question (semantic condition)	Wer angelt alles im Garten? who fishes all in.the garden 'Who all is fishing in the garden?'	Exhaustive	Der Großpapa, die Großmutter, die Mutter und der Junge. the grandpa the grandma and mother and the boy 'The grandpa, the grandma, the mother and the boy'
		Non-exhaustive	Der Großpapa, die Großmutter und der Junge. the grandpa the grandma and the boy 'The grandpa and the boy'
		Singleton	Der Junge. the boy 'the boy'
einige (pragmatic condition)	Nenne einige Menschen, die im Garten angeln. name some people the in.the garden fish 'Name some people who are fishing in the garden.'	Exhaustive	Der Großpapa, die Großmutter, die Mutter und der Junge. the grandpa the grandma and mother and the boy 'The grandpa, the grandma, the mother and the boy'
		Non-exhaustive	Der Großpapa, die Großmutter und der Junge. the grandpa the grandma and the boy 'The grandpa and the boy'
		Singleton	Der Junge. the boy 'the boy'
die (plural definite descriptions, semantic condition)	Nenne die Menschen, die im Garten angeln. name the people that.PL in.the garden fish 'Name the people who are fishing in the garden.'	Exhaustive	Der Großpapa, die Großmutter, die Mutter und der Junge. the grandpa the grandma and mother and the boy 'The grandpa, the grandma, the mother and the boy'
		Non-exhaustive	Der Großpapa, die Großmutter und der Junge. the grandpa the grandma and the boy 'The grandpa and the boy'
		Singleton	Der Junge. the boy 'the boy'

comparisons of all levels of answer type (i.e. singleton, non-exhaustive and exhaustive). Our third modification was that we reduced the number of trials by about 50%, as some respondents in Experiment 1 had remarked on the high number of trials. Parallel to Experiment 1, we explored whether non-exhaustive answers to the bare *wh*-question would pattern with those to *wer alles* and the plural definite description or with *einige*.

3.1. Method

3.1.1. Participants

307 native speakers of German (74 males and 233 females) from the University of Oldenburg, other universities and from the internet (aged from 17 to 70, *Mean* = 25.73, *SD* = 7.41) completed our survey. They were ordered randomly to one of the 12 counterbalance lists of the questionnaire, so that a close to equal number of participants was assigned randomly to every list. Incomplete questionnaires or those filled in by non-native speakers were excluded from the analyses (around 20% of the all questionnaires were discarded as a result of these criteria).

3.1.2. Material

The questionnaire comprised 56 questions that were classified into 36 critical items, 15 filler items and 5 practice items. The critical items and the fillers were adopted from Experiment 1 except that we used pictures with four persons performing the critical action and correspondingly four-person answers in the small dialogues in the exhaustive condition. As in Experiment 1, the questions were followed by answers that the participant had to rate on a scale of 1–6.

In the critical trials, the answers either listed one person out of four (singleton answer), three people of four (non-exhaustive answer) or all four people (exhaustive answer). Our set-up gave rise to 12 experimental conditions. *Wer*, *wer alles*, plural definite descriptions and *einige* were all tested with singleton, non-exhaustive, and exhaustive answers (see Table 3). Figure 3 illustrates a typical picture stimulus.

3.1.3. Procedure

The procedure was the same as in Experiment 1 with the only exception that in this experiment participants filled in the questionnaire on the LimeSurvey platform to allow for complete random presentation of items within participants.

3.1.4. Primary Statistical Analyses: Cumulative Link Mixed Models (CLMM)

Statistical analyses are the same as described in Experiment 1 (see Section 2.1.4).

3.2. Results

Table 4 illustrates the estimated means and medians of scores per Sentence Type and Answer Type with standard deviations (SD), standard errors (SE) and 95% confidence intervals of the mean (CI). Estimations are computed from 921 items (3 items per condition per participant * 307 participants).

Median scores for the exhaustive responses were 6, as in Experiment 1, and the non-exhaustive responses to *wer alles*, and to plural definite descriptions received a median score of 4, whereas *einige* received a median score of 6 and *wer* a median score of 5. Responses in the singleton conditions all received a median score of 2, much lower than the ratings of the non-exhaustive responses. SD inspection based on Table 4 shows that scores associated to infelicitous responses are more spread than the ratings to felicitous

Table 4: Estimated means and medians of all the critical conditions in Experiment 2.[11]

Sentence Type	Answer Type	Mean	Median	SD	SE	95% CI
wer	Exhaustive	5.881	6	0.486	0.016	0.031
	Non-exhaustive	4.325	5	1.025	0.034	0.066
	Singleton	2.926	2	1.416	0.047	0.092
wer alles	Exhaustive	5.879	6	0.567	0.019	0.037
	Non-exhaustive	3.971	4	1.132	0.037	0.073
	Singleton	2.020	2	0.743	0.024	0.048
die[12]	Exhaustive	5.887	6	0.536	0.018	0.035
	Non-exhaustive	4.001	4	1.117	0.037	0.072
	Singleton	2.050	2	0.753	0.025	0.049
einige	Exhaustive	5.682	6	0.724	0.024	0.047
	Non-exhaustive	5.265	6	1.025	0.034	0.066
	Singleton	2.809	2	1.220	0.040	0.079

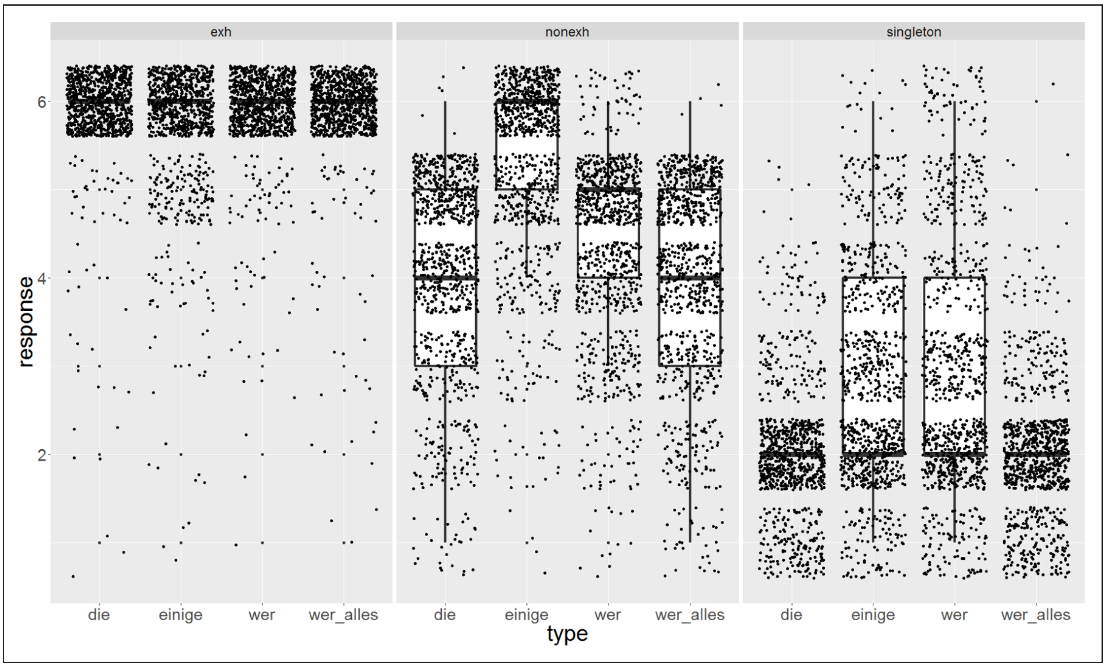

Figure 4: "Jitter" plot with the distribution of responses in the critical conditions in Experiment 2.

responses, just like in Experiment 1. Figure 4 illustrates the distribution of responses visually in the critical Sentence Type conditions as a function of Answer Type.

A CLMM with Sentence Type, Answer Type and their interaction was computed. List as a random factor did not increase explanatory power of the model as revealed by model comparisons. The best-fitting random-structure contained participant and item as random factors.

First, an intercept-only model was built. Then, Answer Type, Sentence Type and their interaction term were entered into CLMM analyses in a stepwise manner. The best-fitting

[11] In Experiment 2 non-exhaustive responses enumerated 3 out of 4 people.
[12] Plural definite descriptions.

model included Answer Type ($p < 0.001$), Sentence Type ($p < 0.001$) and their interaction ($p < 0.001$) as fixed factors, indicating that all these three terms were significant predictors of felicity ratings.

It is noteworthy that the singleton violations in the *wer*-condition are far more spread than those in the condition with plural definite descriptions or the *wer alles*-conditions. This effect in the *wer-singleton* condition is reflected by a very high between-participant variance in the median ratings (*Mean* = 2.94, *Median* = 2, *Mode* = 2, *Variance* = 1.925, $N = 307$ participants). This indicates that the singleton responses were rated very differently between participants, while the mean within-participant SD of ratings was low (1.42). The high between-participant variance reflects that around 20% of the participants consistently accepted the singleton violation as a legitimate answer (median rating of 5 or 6), while half of the respondents rejected it systematically, yielding a median rating of 1 or 2 (for the distribution of median responses in the *wer*-singleton condition, see Appendix C).

For 169 out of 307 participants (55%) the mean difference between the non-exhaustive *wer*-condition and the non-exhaustive condition with plural definite descriptions is smaller than 0.25 (with non-exhaustive responses to *wer* rated as more felicitous), while the group-level mean difference between these conditions is 0.324, comparable to the difference in Experiment 1.

4. Further Statistical Analyses for the two experiments: Estimates of Minimal Important Difference (MID)

Because experimental outcomes are highly dependent on *a priori* factors such as random measurement error, Type-I (false-positive error) and Type-II (false-negative error) statistical errors or sample size inflation (Lin et al. 2013), even a linguistically irrelevant difference can become statistically significant leading to rejection of the null hypothesis (Farivar et al. 2004; Kalinowski & Fidler 2010). Sullivan & Feinn (2012) argue that differences between any sample means will turn significant if the sample is large enough. To avoid this fallacy, we evaluated the magnitude of differences (as suggested by Farivar et al. 2004).

To tease apart statistical significance and linguistic meaningfulness,[13] we calculated Minimal Important Difference measures (MID), first described by Jaeschke et al. (1989), using a standard deviation criterion and effect size parameters in conjunction, according to recent recommendations about the triangulation of these methods (Crosby et al. 2003; Revicki et al. 2006; 2008; Engel et al. 2018).

MID is defined as "the smallest change in score in the construct to be measured which patients perceive as important" (Mokkink et al. 2010).[14] The MID in our context is a measure for the smallest change score of interest which participants perceive as relevant or important. We interpreted the between-conditions criterion of MID as denoting the smallest quantity of interest, i.e., the smallest difference in Likert-scores between two conditions, such as *wer* and *wer alles*, that is linguistically meaningful regarding the nature of the relevant structures. In other words, we expect that if two minimally different structures differ in their linguistic nature, then they will be perceived as relevantly different by participants, as in the Drenhaus et al. (2011) study.

We developed MIDs using so-called distribution-based techniques, which are based on the statistical characteristics of the obtained sample (Crosby et al. 2003). One advantage

[13] Sometimes referred to as "practical significance" or "practical importance" (see, for example, Kalinowski & Fidler 2010).

[14] "Important" and "relevant" are used synonymously in the MID-literature.

of this is that they account for differences beyond some level of random variation (Crosby et al. 2003). According to the most commonly used approach, the one-half standard deviation criterion of an outcome measure indicates that a difference of more than one-half of the outcome score's standard deviation can be regarded as a substantially important difference (Norman et al. 2003; 2004; Copay et al. 2007).

Norman et al. (2003) provide a systematic review of literature on MIDs reported in 38 studies. 32 out of the 38 studies in their review reported MIDs close to half a SD. Norman et al. (2003; 2004) argue that this value is universal and provide psychophysiological evidence for this threshold: "[the] explanation for this consistency is that research in psychology has shown that the limit of people's ability to discriminate over a wide range of tasks is approximately 1 part in 7, which is very close to half a SD (Norman et al. 2003: 582)."

Whereas Norman et al. (2004) argue for the 0.5 SD criterion, Farivar et al. (2004) and Eton et al. (2004) suggest an SD of 0.3. To keep the MID as low as possible, we adopt the more conservative, one-third SD-criterion. We do not know of any studies which have reported an MID below this threshold.

Another way of addressing the risk of overestimating differences has been suggested by Vacha-Haase (2001), Norman (2005), Sullivan & Feinn (2012), Lin et al. (2013: 4), and Vasishth & Gelman (2017) among many others (including APA 2010: 33), who recommend reporting effect sizes to step around the problem of highly significant but meaningless differences due to large sample sizes. A cut-off point of effect size can be used to define MID in the same way as the SD-criteria. In order to do this, we evaluated the magnitude of effects using a distribution-free effect size measure, instead of the often-used parametric index Cohen's *d* (Cohen 1988). Specifically, we used Cliff's delta as a quantitative measure of the strength of an effect (Cliff 1993; 1996), which quantifies the amount of difference between two non-parametric variables. Unlike Cohen's *d*, Cliff's *δ* is bounded, with a δ of 1 or –1 at the two ends. In other words, the δ index represents the degree of overlap between two distributions of ordinal scores, with a δ-value of 0 signaling overlap between the two groups, i.e., an effect size of zero.

In the interpretation of δ, we used the benchmarks provided in Romano et al. (2006): $|\delta| < 0.147$ "negligible", $0.147 \leq |\delta| < 0.33$ "small", $0.33 \leq |\delta| < 0.474$ "medium", else "large" ($|\delta| \geq 0.474$). We set this threshold effect size at the traditional limit of small effect size, 0.33 following a common MID guideline stating that small effect sizes designate an irrelevant difference (Norman et al. 2003; Angst et al. 2017). To compute Cliff's δ effect sizes, we used the R-package "effsize" (Torchiano 2016). Crucially, the use of multiple strategies in determining MIDs strengthens the interpretability of any statistically significant difference between linguistic structures.

4.1. MID estimates in Experiment 1

Table 5 summarizes the MID estimates using the distribution-based techniques including standard deviation criteria and effect sizes, and the significance values of the comparisons. Multiple comparisons between experimental conditions were based on the best-fitting CLMMs.

For non-exhaustive answers, the statistically significant differences between *wer* – *wer so alles*, *wer* – plural definite descriptions and *wer so alles* – plural definite descriptions did not yield linguistically meaningful differences, since they did not reach the MID-threshold (see Nature of difference, Table 5). Effect size statistics corroborated these results.

Table 5: Observed absolute differences and multiple comparisons for the critical contrasts in Experiment 1.[15]

Comparisons	Observed mean difference	adj. *p*-value (comparison)	MID SD-criteria 1/3 SD (Eton et al. 2004)	Cliff's *δ* (95% CI)	Nature of difference
wer – wer so alles (non-exhaustive)	0.204	9.400e-06***	0.3843	0.11 [–0.02; 0.24]	trivial
wer – die (non-exhaustive)	0.366	8.319e-16***	0.3880	0.18 [0.11; 0.24]	trivial
wer so alles – die (non-exhaustive)	0.162	2.772e-04***	0.3843	0.08 [0.01; 0.14]	trivial
wer – einige (non-exhaustive)	1.05	2.644e-84***	0.374	–0.52 [–0.57; –0.46]	non-trivial
einige – die (non-exhaustive)	1.416	1.982e-116***	0.3880	–0.63 [–0.68; –0.58]	non-trivial
einige – wer so alles (non-exhaustive)	1.254	3.627e-101***	0.3843	0.59 [0.53; 0.64]	non-trivial
wer so alles – wer alles (singleton)	0.282	1.412e-10***	0.3977	0.13 [0.06; 0.20]	trivial

In contrast, for non-exhaustive answers the statistically significant differences between *wer – einige*, *einige –* plural definite descriptions and *einige – wer so alles* revealed linguistically meaningful differences. For the singleton answers, the statistically significant difference between *wer so alles – wer alles* again was not linguistically meaningful.

4.2. MID estimates in Experiment 2

Table 6 summarizes the MID estimates for Experiment 2 in the same way as for Experiment 1. For the non-exhaustive answers, the statistically significant differences between *wer – wer alles* and *wer – die* (plural definite descriptions) did not yield linguistically meaningful differences, since they did not reach the threshold for MID by the one-third SD criterion. Also, Cliff's *δ* effect sizes of –017 (*wer-wer alles*) and 0.16 (*wer-die*) for non-exhaustive answers are comparable to the Cliff's *δ* effect size of 0.18 for the *wer-die* comparison in Experiment 1 (non-exhaustive answer).

For non-exhaustive answers, the difference between *wer alles – die* did not even show a statistically significant difference. For non-exhaustive answers, the statistically significant differences between *wer – einige*, *einige – die* and *einige – wer alles* all revealed linguistically meaningful differences. For the singleton answer, the statistically significant differences between *wer alles – einige*, *wer – wer alles* and *wer – die* proved to be linguistically meaningful. For the singleton answer, the differences between the comparisons *einige – wer* and *wer alles – die* were not significant.

[15] Significance codes: *** *p*-value < 0.001, ** *p*-value < 0.01, * *p*-value < 0.05. Multiple comparisons were carried out applying the Benjamini & Hochberg adjustment (1995). Estimated Minimal Important Differences (MID) are based on the one-third SD criterion and effect size estimates (Cliff's δ) with 95% confidence intervals (CI) of the effect sizes. Interpretations of Cliff's δ according to Romano et al. (2006): $|\delta| < 0.147$ "negligible", $0.147 \leq |\delta| < 0.33$ "small", $0.33 \leq |\delta| < 0.474$ "medium", else "large" ($|\delta| \geq 0.474$). Negligible and small effect sizes signal trivial differences, while medium or large effect sizes meaningful differences indicating theoretical significance. To convert the scientific numbers to decimal numbers, move the decimal point n places to the left, as indicated by the –n located after the "e". The condition with plural definite descriptions is referred to as *die*.

Table 6: Observed absolute differences and multiple comparisons for the critical contrasts in Experiment 2.

Comparisons	Observed mean difference	adj. *p*-value (comparison)	MID SD-criteria 1/3 SD (Eton et al. 2004)	Cliff's *δ* (95% CI)	Nature of difference
wer – wer alles (non-exhaustive)	0.354	2.578e-21***	0.377	−0.17 [−0.21; −0.12]	trivial
wer – die (non-exhaustive)	0.324	6.233e-18***	0.372	0.16 [0.11; 0.21]	trivial
wer alles – die (non-exhaustive)	0.03	0.2665 n.s.	0.372	−0.01 [−0.06; 0.04]	no difference
wer – einige (non-exhaustive)	0.94	3.306e-142***	0.342	−0.53 [−0.57; −0.49]	non-trivial
einige – die (non-exhaustive)	1.264	6.151e-179***	0.372	−0.63 [−0.67; −0.59]	non-trivial
einige – wer alles (non-exhaustive)	1.294	1.901e-187***	0.342	−0.64 [−0.67; −0.60]	non-trivial
wer alles – einige (singleton)	0.789	1.056e-94***	0.407	−0.37 [−0.42; −0.33]	non-trivial
wer alles – die (singleton)	0.03	0.1659 n.s.	0.251	−0.02 [−0.06; 0.03]	no difference
wer – wer alles (singleton)	0.906	1.264e-80***	0.472	−0.37 [−0.41; −0.33]	non-trivial
einige – wer (singleton)	0.117	0.06165 n.s.	0.406	−0.02 [−0.07; 0.03]	no difference
wer – die (singleton)	0.876	2.661e-79***	0.472	0.37 [0.33; 0.41]	non-trivial

5. Discussion

In two experiments, felicity judgment data were analyzed to examine the extent of exhaustivity violations in five sentence type conditions comparing *wer*, *wer alles*, and *wer so alles* with plural definite descriptions and with *einige*. We addressed the following question: Do bare *wh*-questions (*wer*) pattern with the *wer alles*-structure, with plural definite descriptions triggered by the determiner *die*, or with *einige* ('some')? Based on Schulz & Roeper (2011) we explored whether the judgment of answers to *wer* patterns with that to *wer alles* and plural definite descriptions, which do not allow for non-exhaustive answers, or with *einige*, which can prompt a 'some but not all' response, in a fashion similar to *wer*.

Because of possible task effects and the large sample-size, we applied additional distribution-based analyses, which allowed us to tease apart statistically significant but irrelevant results from statistically significant and linguistically meaningful results. We operationalized the question of whether observed differences between structures are big enough to reach linguistic meaningfulness in the following way: the magnitude of an observed difference required surpassing the MID-threshold for linguistic meaningfulness and differences smaller than the MID were labelled as trivial, i.e. linguistically irrelevant.

Before we turn to the results in detail, we need to address one general finding and one unexpected result. The general finding concerning all structures was that singleton answers were rated lower than non-exhaustive answers. This pattern can be explained by two effects: participants may have graded the incomplete answers in terms of their quantity of truth-content, and the Likert-scale invites dichotomous response behaviour (on context effects on respondents' answers, see, for example, Schwarz & Strack 1990). The unexpected result concerns the exhaustive answers to *einige* that in both experiments received a higher acceptance rate than the non-exhaustive answers. It may be due to the specific set-up of the task. Recall that all pictures contained several individuals performing the relevant action and some individuals performing contrasting actions. Therefore, it seems likely that the pictures invited a contrastive reading of the *einige*-statement. Exhaustive answers, mentioning all individuals performing the relevant action, were probably judged as felicitous, because those individuals who performed another action were correctly excluded from this list answer. For example, in a picture with four people fishing, one ironing, and one taking a photo, participants judged the "exhaustive" answer enlisting the four people fishing as more felicitous than the one mentioning just three. A further study could investigate the role of the presence of individuals who perform a different action on the judgment of non-exhaustive responses in the *einige*-condition. As it is, the presence of these individuals may have influenced the judgments.

This result in the *einige*-condition could also be due to the ratio between the overall set-size and the relevant sub-set. Degen & Tanenhaus (2011; 2015), for example, showed that *some* is less natural for reference to small compared to intermediate sub-sets, and that *some* referring to smaller set-sizes is more readily interpreted as 'possibly all'. Finally, the school-context of the task may have prompted participants to accept exhaustive answers more readily, as informativeness in students' answers is usually rewarded.

In the sections to follow, we discuss the findings from the four central comparisons that were carried out to examine the linguistic relationship between the respective structures: *wer alles* versus plural definite descriptions (Section 5.1), *wer alles* versus *einige* (Section 5.2), *wer* versus *wer alles* versus plural definite descriptions (Section 5.3), and *wer* versus *einige* (Section 5.4). The last two comparisons are crucial to determine the linguistic nature of exhaustivity in bare *wh*-questions. Parallels with plural definite descriptions, and differences to *einige,* would favor a semantic analysis of bare *wh*-questions.

5.1. Wer alles versus plural definite description

Judgments in the conditions *wer alles* and plural definite descriptions were found to be comparable despite their differences in form. In Experiment 2, *wer alles* and plural definite descriptions did not show any differences, neither statistically nor regarding the MID, in judgments of exhaustive, non-exhaustive and even singleton answers, providing empirical evidence for the assumption that the two structures share crucial semantic properties.

5.2. Wer alles versus einige

Judgments in the conditions *wer alles* and *einige* were different for both singleton and non-exhaustive answers in Experiment 2. We found non-trivial, i.e., linguistically meaningful, differences in both singleton and non-exhaustive conditions: a singleton answer to *wer alles* was considerably less acceptable than a singleton answer to *einige*. Likewise, a non-exhaustive answer to *wer alles* was considerably less acceptable than a non-exhaustive answer to *einige*. This indicates a difference in linguistic nature in these two structures.

5.3. Wer versus wer alles versus plural definite descriptions

In Experiment 1, judgments of non-exhaustive answers to *wer* patterned with those to plural definite descriptions. In Experiment 2, judgments of non-exhaustive answers to *wer* patterned with those to plural definite descriptions and with *wer alles*. The small magnitude of differences between *wer*, *wer alles*, and plural definite descriptions in non-exhaustive conditions do not exceed the MID-thresholds in either of the experiments.[16]

The small magnitude of differences is driven by inter-individual variance in interpretation rather than by systematic within-participant differences which would have caused a reliable group-level difference. Only 20% of the participants in Experiment 1, and 55% of the participants in Experiment 2[17] distinguished between non-exhaustive answers to *wer* (more felicitous) and non-exhaustive answers to plural definite descriptions (less felicitous), pointing to inter-individual variation.

Based on these results, we argue that exhaustivity triggered by plural definite descriptions contained in a restrictive relative clause is similar to the exhaustivity requirement of bare *wh*-questions in "mention-all" contexts. We take these findings to support the semantic analysis of exhaustivity in bare *wh*-questions as suggested by Nishigauchi (1999) and Schulz & Roeper (2011). In this line of argumentation, bare *wh*-questions in the "mention-all" reading can be assumed to contain an underlying universal quantifier in a parallel fashion to plural definite descriptions contained in a restrictive relative clause. Note that given the limits of the offline-judgment task, we cannot rule out that some participants interpreted the *wh*-questions as a "mention-some" request and may have arrived at similar ratings as those participants who interpreted the *wh*-questions as "mention-all".[18]

Unlike for non-exhaustive answers, for singleton answers to *wer* vs. *wer alles* and *wer* vs. plural definite descriptions we found a linguistically meaningful difference of around 0.9. We argue that the difference between *wer* and *wer alles* in the singleton condition is caused by the fact that the lexical marker *alles* encodes the exhaustivity requirement lexically, only allowing "mention-all" readings. As *wer* is ambiguous between "mention-some" and "mention-all" readings, a singleton answer may reflect a "mention-some" reading and hence result in a higher acceptability. Importantly, 50% of the participants showed a median response of 1 or 2 in this condition, indicating that they interpreted this structure as being equal to *wer alles*.

5.4. *Wer versus einige*

As expected, non-exhaustive answers to *einige* were accepted to a higher degree than the non-exhaustive answers to bare *wh*-questions in both experiments. This difference was non-trivial and indicates that the same set-size of individuals exerts a different effect on the perceived felicity of the answers to *einige* and to bare *wh*-questions. As already mentioned in Section 5, the inclusion of contrastive actions might have influenced the acceptability of either of the response types in the case of *einige*.

Singleton answers to *einige* and *wer* did not reveal any difference, with ratings being equally low. According to the semantic analysis singleton answers to *wer* should be dispreferred, which was supported by our results. Does this mean that *einige* received a semantic

[16] Small differences between *wer* and *wer alles* below the MID may be due to the presence of lexical alternatives within the same experiment: the *wer alles*-sentences may have invited participants to contrast *wer* with *wer alles* and to rate *wer* differently from *wer alles*. See Degen & Tanenhaus (2015) who report that the availability of lexical alternatives causes a pragmatic effect.

[17] The difference between the two experiments may be due to the size of the relevant sub-set of individuals performing the action: 2 out of 3 (Experiment 1) versus 3 out of 4 (Experiment 2). Condition-means and condition-medians are in line with this difference between the two experiments for the non-exhaustive *wer*-condition: mean = 4.062, median = 4 in Experiment 1, and mean = 4.325, median = 5 in Experiment 2.

[18] As pointed out by one of the reviewers, we cannot infer from the raw ratings *per se* whether participants interpreted the *wh*-questions exhaustively or non-exhaustively. For example, a rating of 4 may reflect the subjective evaluation of an incomplete answer according to the participant's exhaustive interpretation (on the exhaustive reading, incomplete answers are "incorrect"), or may indicate the subjective evaluation of the same incomplete answer according to the participant's non-exhaustive reading (on the non-exhaustive reading, incomplete answers are "correct"). For the purposes of this paper we assume the stronger hypothesis that ratings reflect the subjective evaluation of exhaustivity violations in the case of incomplete answers in the *wer*-condition (violation of "mention-all").

interpretation? We suggest that the reason for the similar low rating for singleton answers to *einige* is not the same as for *wer*: the singleton answer to *einige* simply represents a violation of number because *einige Leute/Menschen* ('some people') denotes plural, and this is what caused the low ratings.

5.5. Summary of discussion

Evaluating the data for plural definite descriptions, *wer* and *wer alles* in non-exhaustive conditions together, we found the magnitude of pair-wise differences between these three structures in the non-exhaustive conditions to be below the Minimal Important Difference limit (MID) in all cases. Crucially, the differences do not pass the strictest one-third SD-criterion for MID (Eton et al. 2004; Farivar et al. 2004) and do not approach the medium threshold for effect size defined by Romano et al. (2006), which suggests that these structures are alike. These parallels between *wer* and plural definite descriptions and *wer alles* are expected under a semantic account. Notably, our findings are consistent with recent research in language acquisition. A comprehension study with four-to-six-year-old children did not find any relationship between the mastery of pragmatics of quantification using scalar implicature and mastery of exhaustivity in *wh*-questions, suggesting that the two are not derived by a common pragmatic process (Foryś et al. 2017).

Turning to the comparison with the scalar quantifier *einige*, the difference we observed between *wer* and *einige* is in line with a semantic account as well. We acknowledge that an alternative explanation exists according to which *einige* and *wer* have a similar meaning representation; the differences in mean ratings would then result from the differences in the effect of the determining context on the requirements to exhaust in these conditions. It is not clear to us at this point how this explanation could account for all our data in the same way as the semantic account does. Therefore, given the clear parallels between *wer* and plural definite descriptions and *wer alles* we conclude that exhaustivity in *wh*-questions is of semantic nature.

6. Conclusion

The objective of our two experiments was to assess the judgment of incomplete answers to unembedded bare *wh*-questions in German (*wer* 'who') in comparison to *wer alles*-questions ('who all'), plural definite descriptions containing a restrictive relative clause (*die Leute, die* 'the people who') and structures containing "some" (*einige*). To this end, we developed a felicity judgment experiment, modelled after the Exhaustivity task (see Schulz & Roeper 2011; Schulz 2015), testing a total of 441 monolingual speakers of German. A novel way of using Minimal Importance Difference (MID) estimates was employed, for expression of experimental differences by means of the MID provides a more careful insight into the nature of differences between linguistic phenomena than traditional significance testing.

Specifically, using MID based on the SD- and effect size criteria we examined whether exhaustivity violations of bare *wh*-questions pattern with that of *wer alles* and with plural definite descriptions or with the infelicitous answers to *einige*. Similarity with *wer alles* and plural definite descriptions was argued to provide evidence for a semantic analysis of bare *wh*-questions, whereas similarity with *einige* was argued to point to a pragmatic analysis. We found that unlike singleton answers, non-exhaustive answers to *wer*, to *wer alles* and to plural definite descriptions patterned together. Furthermore, non-exhaustive answers to *einige* were accepted to a higher degree than non-exhaustive answers to *wer*. These findings point to the semantic nature of exhaustivity in bare *who*-questions in line with Nishigauchi (1999), Nelken & Shan (2004), and Schulz & Roeper (2011).

To our knowledge, this study is the first to report MID estimates for exhaustivity violations, and therefore could inspire future studies aiming to derive MIDs across various types of violations for different phenomena. In future research the MID should be embedded in the weight of evidence together with traditional statistical comparisons.

Abbreviations

ACC = Accusative, AIC = Akaike Information Criterion, CI = Confidence Interval, CLMM = Cumulative Link Mixed Models, DP = Determiner Phrase, EEG = Electroencephalography, ERP = Event-Related Potentials, FDR = False Discovery Rate, IMP = Imperative, MID = Minimal Important Difference, NOM = Nominative, PL = Plural, SD = Standard Deviation, SE = Standard Error

Additional Files

The additional files for this article can be found as follows:

- **Appendix A.** Instructions in Experiments 1 and 2. DOI: https://doi.org/10.5334/gjgl.549.s1
- **Appendix B.** Critical and filler items in Experiments 1 and 2. DOI: https://doi.org/10.5334/gjgl.549.s1
- **Appendix C.** Distribution of the median responses in the wer-singleton condition in Experiment 2. DOI: https://doi.org/10.5334/gjgl.549.s1

Acknowledgements

The authors thank Cornelia Hamann and to Malte Zimmermann for their helpful comments. They are grateful to the students of the University of Oldenburg for participating in the study. The comments of the three reviewers were also very much appreciated and further improved this paper.

Competing Interests

The authors have no competing interests to declare.

References

Agresti, Alan. 2002. *Categorical data analysis* (2nd ed.). New York, NY: John Wiley & Sons. DOI: https://doi.org/10.1002/0471249688

Akaike, Hirotugu. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6). 716–723. DOI: https://doi.org/10.1109/TAC.1974.1100705

American Psychological Association (APA). 2010. *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Angst, Felix, André Aeschlimann & Jules Angst. 2017. The minimal clinically important difference (MCID) raised the significance of outcome effects above the statistical level, with methodological implications for future studies. *Journal of Clinical Epidemiology* 82. 128–136. DOI: https://doi.org/10.1016/j.jclinepi.2016.11.016

Antoniou, Kyriakos, Chris Cummins & Napoleon Katsos. 2016. Why only some adults reject under informative utterances. *Journal of Pragmatics* 99. 78–95. DOI: https://doi.org/10.1016/j.pragma.2016.05.001

Barbet, Cécile & Guillaume Thierry. 2016. Some alternatives? Event-related potential investigation of literal and pragmatic interpretations of *some* presented in isolation. *Frontiers in Psychology* 7(403). 1479. DOI: https://doi.org/10.3389/fpsyg.2016.01479

Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3). 255–278. DOI: https://doi.org/10.1016/j.jml.2012.11.001

Beck, Sigrid & Hotze Rullmann. 1999. A flexible approach to exhaustivity in questions. *Natural Language Semantics* 7. 249–298. DOI: https://doi.org/10.1023/A:100837322

Benjamini, Yoav & Yosi Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* 57. 289–300.

Brouwer, Harm, Matthew W. Crocker, Noortje J. Venhuizen & John C. J. Hoeks. 2017. A neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science* 41(S6). 1318–1352. DOI: https://doi.org/10.1111/cogs.12461

Caponigro, Ivano, Lisa Pearl, Neon Brooks & David Barner. 2012. Acquiring the meaning of free relative clauses and plural definite descriptions. *Journal of Semantics* 29. 261–293. DOI: https://doi.org/10.1093/jos/ffr014

Chierchia, Gennaro, Danny Fox & Benjamin Spector. 2012. Scalar implicature as a grammatical phenomenon. In Claudia Maienborn, Klaus von Heusinger & Paul Portner (eds.), *Semantics: An International Handbook of Natural Language Meaning* 3. 2297–2331. Berlin: Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110253382.2297

Christensen, Rune Haubo B. 2015. Ordinal – Regression models for ordinal data. R package version 2015, 6–28. http://www.cran.r-project.org/package = ordinal/.

Cliff, Norman. 1993. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin* 114. 494–509. DOI: https://doi.org/10.1037//0033-2909.114.3.494

Cliff, Norman. 1996. *Ordinal methods for behavioral data analysis*. Mahwah, NJ: Lawrence Erlbaum.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences* (2nd ed.). NJ: Lawrence Erlbaum Associates. DOI: https://doi.org/10.4324/9780203771587

Copay, Anne G., Brian R. Subach, Steven D. Glassman, David W. Polly & Thomas C. Schuler. 2007. Understanding the minimum clinically important difference: A review of concepts and methods. *The Spine Journal* 7. 541–546. DOI: https://doi.org/10.1016/j.spinee.2007.01.008

Crosby, Ross D., Ronette L. Kolotkin & G. Rhys Williams. 2003. Defining clinically meaningful change in health-related quality of life. *Journal of Clinical Epidemiology* 56. 395–407. DOI: https://doi.org/10.1016/S0895-4356(03)00044-1

Cummings, Louise. 2015. Theory of mind in utterance interpretation: The case from clinical pragmatics. *Frontiers in Psychology* 6. 1286. DOI: https://doi.org/10.3389/fpsyg.2015.01286

Degen, Judith & Michael K. Tanenhaus. 2011. Making inferences: The case of scalar implicature processing. In Laura Carlson, Christoph Hölscher & Thomas F. Shipley (eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 3299–3304.

Degen, Judith & Michael K. Tanenhaus. 2015. Processing scalar implicature: A constraint-based approach. *Cognitive Science* 39(4). 667–710. DOI: https://doi.org/10.1111/cogs.12171

Degen, Judith & Noah D. Goodman. 2014. Lost your marbles? The puzzle of dependent measures in experimental pragmatics. In Paul Bello, Marcello Guarini, Marjorie McShane & Brian Scassellati (eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 397–402.

de Villiers, Jill G. & Tom Roeper. 1993. The emergence of bound variable structures. In Werner Abraham & Eric Reuland (eds.), *Knowledge and Language: Orwell's problem and Plato's problem*, 105–139. Dordrecht: Kluwer.

Dieussaert, Kristien, Suzanne Verkerk, Ellen Gillard & Walter Schaeken. 2011. Some effort for some: Further evidence that scalar implicatures are effortful. *The Quaterly Journal of Experimental Psychology* 64(12). 2352–2367. DOI: https://doi.org/10.1080/17470218.2011.588799

Drenhaus, Heiner, Douglas Saddy & Stefan Frisch. 2005. Processing negative polarity items: When negation comes through the backdoor. In Stephan Kepser & Marga Reis (eds.), *Linguistic evidence: Empirical, theoretical and computational perspectives*, 145–165. Berlin: Mouton de Gruyter.

Drenhaus, Heiner, Malte Zimmermann & Shravan Vasishth. 2011. Exhaustiveness effects in clefts are not truth-functional. *Journal of Neurolinguistics* 24(3). 320–337. DOI: https://doi.org/10.1016/j.jneuroling.2010.10.004

Drenhaus, Heiner, Peter beim Graben, Douglas Saddy & Stefan Frisch. 2006. Diagnosis and repair of negative polarity constructions in the light of symbolic resonance analysis. *Brain and language*, 255–268. DOI: https://doi.org/10.1016/j.bandl.2005.05.001

Dupuy, Ludivine, Penka Stateva, Sara Andreetta, Anne Cheylus, Viviane Déprez, Jean-Baptiste van der Henst, Jacques Jayez, Arthur Stepanov & Anne Reboul. 2018. Pragmatic abilities in bilinguals: The case of scalar implicatures. *Linguistic Approaches to Bilingualism*. DOI: https://doi.org/10.1075/lab.17017.dup

Engel, Lisa, Dorcas E. Beaton & Zahi Touma. 2018. Minimal clinically important difference: A review of outcome measure score interpretation. *Rheumatic Disease Clinics of North America* 44(2). 177–188. DOI: https://doi.org/10.1016/j.rdc.2018.01.011

Eton, David T., David Cella, Kathleen J. Yost, Susan E. Yount, Amy H. Peterman, Donna S. Neuberg, George W. Sledge & William C. Wood. 2004. A combination of distribution-and anchor based approaches determined minimally important differences (MIDs) for four endpoints in a breast cancer scale. *Journal of Clinical Epidemiology* 57. 898–910. DOI: https://doi.org/10.1016/j.jclinepi.2004.01.012

Farivar, Sepideh S., Honghu Liu & Ronald D. Hays. 2004. Half standard deviation estimate of the minimally important difference in HRQOL scores? *Expert Review of Pharmacoeconomics & Outcomes Research* 4(5). 515–523. DOI: https://doi.org/10.1586/14737167.4.5.515

Fischler, Ira, Paul A. Bloom, Donald G. Childers, Salim E. Roucos & Nathan W. Perry, Jr. 1983. Brain potentials related to stages of sentence verification. *Psychophysiology* 20. 400–409. DOI: https://doi.org/10.1111/j.1469-8986.1983.tb00920.x

Fodor, Janet D. 1998. Learning to parse? *Journal of Psycholinguistic Research* 27(2). 285–319. DOI: https://doi.org/10.1023/A:1023258301588

Fodor, Janet D. 2002. Prosodic disambiguation in silent reading. In Masako Hirotani (ed.), *Proceedings of the North East Linguistic Society 32*, 113–132. Amherst, MA.

Foryś, Małgorzata, Ewa Haman, Napoleon Katsos & Petra Schulz. 2017. Exploring syntactic, semantic and pragmatic correlates of the acquisition of exhaustivity in wh-questions: A study of Polish monolingual children. *Language Acquisition* 24(1). 27–51. DOI: https://doi.org/10.1080/10489223.2016.1179744

Friederici, Angela D. & Jürgen Weissenborn. 2007. Mapping sentence form onto meaning: the syntax-semantic interface. *Brain Research* 1146. 50–58. DOI: https://doi.org/10.1016/j.brainres.2006.08.038

Gerőcs, Mátyás, Anna Babarczy & Balázs Surányi. 2014. Exhaustivity in focus: Experimental evidence from Hungarian. In Joseph Emonds & Markéta Janebová (eds.), *Language Use and Linguistic Structure. Olomouc Modern Language Series* 3. 181–194. Olomouc: Palacký University.

Gouvea, Ana C., Colin Phillips, Nina Kazanina & David Poeppel. 2010. The linguistic processes underlying the P600. *Language and Cognitive Processes* 25(2). 149–188. DOI: https://doi.org/10.1080/01690960902965951

Grice, Herbert Paul. 1975. Logic and conversation. In Peter Cole & Jerry L. Morgan (eds.), *Syntax and semantics* 3. 41–58. New York, NY: Academic Press.

Grice, Herbert Paul. 1989. *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Groenendijk, Jeroen & Martin Stokhof. 1984. Studies on the semantics of questions and the pragmatics of answers. University of Amsterdamdissertation.

Guasti, M. Teresa, Gennaro Chierchia, Stephen Crain, Francesca Foppolo, Andrea Gualmini & Luisa Meroni. 2005. Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes* 20(5). 667–696. DOI: https://doi.org/10.1080/01690960444000250

Hagoort, Peter, Colin M. Brown & Lee Osterhout. 1999. The neurocognition of syntactic processing. In Colin M. Brown & Peter Hagoort (eds.), *The neurocognition of language*, 273–317. Oxford: Oxford University Press.

Hartshorne, Joshua K., Jesse Snedeker, Stephanie Yen-Mun Liem Azar & Albert E. Kim. 2015. The neural computation of scalar implicature. *Language, Cognition and Neuroscience* 30. 620–634. DOI: https://doi.org/10.1080/23273798.2014.981195

Heim, Irene & Angelika Kratzer. 1998. *Semantics in generative grammar*. Malden, Mass.: Blackwell.

Horn, Laurence. 1972. On the semantic properties of logical operators in English. UCLA, CA dissertation.

Jaeschke, Roman, Joel Singer & Gordon H. Guyatt. 1989. Measurement of health nature: Ascertaining the minimal clinically important difference. *Controlled Clinical Trials* 10(4). 407–415. DOI: https://doi.org/10.1016/0197-2456(89)90005-6

Kaan, Edith. 2007. Event-related potentials and language processing. A brief introduction. *Language and Linguistics Compass* 1(6). 571–591. DOI: https://doi.org/10.1111/j.1749-818X.2007.00037.x

Kalinowski, Pawel & Fiona Fidler. 2010. Interpreting significance: The differences between statistical significance, effect size, and practical importance. *Newborn and Infant Nursing Reviews* 10(1). 50–54. DOI: https://doi.org/10.1053/j.nainr.2009.12.007

Karttunen, Laurie. 1977. Syntax and semantics of questions. *Linguistics and Philosophy* 1. 3–44. DOI: https://doi.org/10.1007/BF00351935

Kasher, Asa, Gila Batori, Nachum Soroker, David C. Graves & Eran Zaidel. 1999. Effects of right and left hemisphere damage on understanding conversational implicatures. *Brain and Language* 68(3). 566–590. DOI: https://doi.org/10.1006/brln.1999.2129

Klinedinst, Nathan & Daniel Rothschild. 2011. Exhaustivity in questions with non-factives. *Semantics and Pragmatics* 4(2). 1–23. DOI: https://doi.org/10.3765/sp.4.2

Kromrey, Jeffrey D. & Kristine Y. Hogarty. 1998. Analysis options for testing group differences on ordered categorical variables: An empirical investigation of Type I error control and statistical power. *Multiple Linear Regression Viewpoints* 25. 70–82.

Kutas, Marta & Steven A. Hillyard. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science* 207(4427). 203–205. DOI: https://doi.org/10.1126/science.7350657

Lau, Ellen F., Colin Phillips & David Poeppel. 2008. A cortical network for semantics: (De)constructing the N400. *Nature Reviews Neuroscience* 9. 920–933. DOI: https://doi.org/10.1038/nrn2532

Lin, Mingfeng, Henry C. Lucas & Galit Shmueli. 2013. Too big to fail: Large samples and the p-value problem. *Information Systems Research* 24(4). 906–917. DOI: https://doi.org/10.1287/isre.2013.0480

Link, Godehard. 1983. The logical analysis of plural and mass terms: A lattice-theoretical approach. In Rainer Bäuerle, Christopher Schwarze & Arnim von Stechow (eds.), *Meaning, use, and interpretation of language*, 302–323. Berlin: de Gruyter. DOI: https://doi.org/10.1515/9783110852820.302

Malamud, Sophia A. 2012. The meaning of plural definites: A decision-theoretic approach. *Semantics and Pragmatics* 5(3). 1–58. DOI: https://doi.org/10.3765/sp.5.3

Mangiafico, Salvatore S. 2015. An R companion for the handbook of biological statistics, version 1.3.2. rcompanion.org/rcompanion/. (Pdf version: http://rcompanion.org/documents/RCompanionBioStatistics.pdf).

Marty, Paul P. & Emmanuel Chemla. 2013. Scalar implicatures: Working memory and a comparison with *only*. *Frontiers in Psychology* 4(403). DOI: https://doi.org/10.3389/fpsyg.2013.00403

Miller, David, David Giancaspro, Mike Iverson, Jason Rothman & Roumyana Slabakova. 2015. Not just algunos, but indeed unos L2ers can acquire scalar implicatures in L2 Spanish. In Anahí Alba de la Fuente, Elena Valenzuela & Cristina Martínez Sanz (eds.), *Language acquisition beyond parameters: Studies in honour of Juana M. Liceras*, 125–145. John Benjamins. DOI: https://doi.org/10.1075/sibil.51.06mil

Mokkink, Lidwine B., Caroline B. Terwee, Donald L. Patrick, Jordi Alonso, Paul W. Stratford, Dirk L. Knol, Lex M. Bouter & Henrica C. W. de Vet. 2010. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health nature measurement instruments: An international Delphi study. *Quality of Life Research* 19. 539–549. DOI: https://doi.org/10.1007/s11136-010-9606-8

Montague, Richard. 1973. The proper treatment of quantification in English. In Jaakko Hintikka, Julius Moravcsik & Patrick Suppes (eds.), *Approaches to natural language*, 221–242. Dordrecht: Reidel.

Nelken, Rani & Chung-chieh Shan. 2004. A Logic of interrogation should be internalized in a modal logic for knowledge. In Kazuha Watanabe & Robert B. Young (eds.), *Proceedings of Semantics and Linguistic Theory [SALT XIV]* 14. 197–211. Ithaca: Cornell University Press. DOI: https://doi.org/10.3765/salt.v14i0.2918

Nieuwland, Mante S., Tali Ditman & Gina R. Kuperberg. 2010. On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language* 63(3). 324–346. DOI: https://doi.org/10.1016/j.jml.2010.06.005

Nishigauchi, Taisuke. 1999. Some preliminary thoughts on the acquisition of the syntax and semantics of wh-constructions. *Theoretical and Applied Linguistics at Kobe Shoin* 2. 35–48.

Norman, Geoffrey R. 2005. The relation between the minimally important difference and patient benefit. *COPD: Journal of Chronic Obstructive Pulmonary Disease* 2(1). 69–73. DOI: https://doi.org/10.1081/COPD-200051249

Norman, Geoffrey R., Jeff A. Sloan & Kathleen W. Wyrwich. 2003. Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care* 41(5). 582–592. DOI: https://doi.org/10.1097/00005650-200305000-00004

Norman, Geoffrey R., Jeff A. Sloan & Kathleen W. Wyrwich. 2004. The truly remarkable universality of half a standard deviation: Confirmation through another look. *Expert*

Review of Pharmacoeconomics & Outcomes Research 4(5). 581–585. DOI: https://doi.org/10.1586/14737167.4.5.581

Osterhout, Lee & Phillip J. Holcomb. 1992. Event related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language* 31. 785–786. DOI: https://doi.org/10.1016/0749-596X(92)90039-Z

Papafragou, Anna & Julien Musolino. 2003. Scalar implicatures: Experiments at the semantics- pragmatics interface. *Cognition* 86. 253–82. DOI: https://doi.org/10.1016/S0010-0277(02)00179-8

Partee, Barbara. 1975. Montague grammar and transformational grammar. *Linguistic Inquiry* 6(2). 203–300.

Pastor-Cerezuela, Gemma, Juan C. Tordera Yllescas, Francisco González-Sala, Maite Montagut-Asunción & María-Inmaculada Fernández-Andrés. 2018. Comprehension of generalized conversational implicatures by children with and without Autism Spectrum Disorder. *Frontiers in Psychology* 9. 272. DOI: https://doi.org/10.3389/fpsyg.2018.00272

Pinheiro, José C. & Douglas M. Bates. 1995. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* 4. 12–35.

Politzer-Ahles, Stephen, Robert Fiorentino, Xiaoming Jiang & Xiaolin Zhou. 2013. Distinct neural correlates for pragmatic and semantic meaning processing: An event-related potential investigation of scalar implicature processing using picture-sentence verification. *Brain Research* 1490. 134–152. DOI: https://doi.org/10.1016/j.brainres.2012.10.042

R Development Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.r-project.org/.

Reich, Ingo. 1997. *Wer will wann wieviel wissen? – Eine Untersuchung verschiedener Frage-Antwort-Bedingungen im Deutschen*. Tübingen/Stuttgart Arbeitspapiere.

Revicki, Dennis A., David Cella, Ron D. Hays, Jeff A. Sloan, William R. Lenderking & Neil K. Aaronson. 2006. Responsiveness and minimal important differences for patient reported outcomes. *Health and Quality of Life Outcomes* 4(70). 1–5. DOI: https://doi.org/10.1186/1477-7525-4-70

Revicki, Dennis A., Ron D. Hays, David Cella & Jeff Sloan. 2008. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology* 61(2). 102–109. DOI: https://doi.org/10.1016/j.jclinepi.2007.03.012

Roeper, Tom & Jill de Villiers. 1991. The emergence of bound variable structures. In Thomas Maxfield, Bernadette Plunkett (eds.), *Papers in the Acquisition of WH. Proceedings of the University of Massachusetts Roundtable (UMOP) GLSA Publications*, 225–266. Amherst, MA: GLSA Publications.

Roeper, Tom, Petra Schulz, Barbara Z. Pearson & Ina Reckling. 2007. From singleton to exhaustive: The acquisition of wh. In Michael Becker & Andrew McKenzie (eds.), *Proceedings of Semantics of Under-Represented Languages in the Americas (SULA)* 3. 87–102. Amherst, MA: GLSA Publications.

Romano, Jeanine, Jeffrey D. Kromrey, Jesse Coraggio & Jeff Skowronek. 2006. Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen's d for evaluating group differences on the NSSE and other surveys? *Paper presented at the Annual meeting of the Florida Association of Institutional Research*. Cocoa Beach, FL.

Romero, Maribel. 2015. Surprise-predicates, strong exhaustivity and alternative questions. In Sarah D'Antonio, Mary Moroney & Carol Rose Little (eds.), *Proceedings of Semantics and Linguistic Theory* 25. 225–245. DOI: https://doi.org/10.3765/salt.v25i0.3081

RStudio. 2012. RStudio: Integrated development environment for R (Version 0.99.892). Boston, MA.

Russell, Benjamin. 2006. Against grammatical computation of scalar implicatures. *Journal of Semantics* 23. 361–382. DOI: https://doi.org/10.1093/jos/ffl008

Sauerland, Uli, Jan Anderssen & Kazuko Yatsushiro. 2005. The plural is semantically unmarked. In Stephan Kepser & Marga Reis (eds.), *Linguistic evidence-empirical, theoretical and computational perspectives*, 409–430. Berlin: Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110197549.413

Schulz, Katrin & Robert van Rooij. 2006. Pragmatic meaning and non-monotonic reasoning: The case of exhaustive interpretation. *Linguistics and Philosophy* 29(2). 205–250. DOI: https://doi.org/10.1007/s10988-005-3760-4

Schulz, Petra. 2015. Exhaustivity. In Sharon Armon-Lotem, Jan de Jong & Natalia Meir (eds.), *Methods for assessing multilingual children: Disentangling bilingualism from Specific Language Impairment*. Clevedon: Multilingual Matters.

Schulz, Petra & Tom Roeper. 2011. Acquisition of exhaustivity in wh-questions: A semantic dimension of SLI? *Lingua* 121. 383–407. DOI: https://doi.org/10.1016/j.lingua.2010.10.005

Schwarz, Florian. 2013. Maximality and definite plurals – experimental evidence. *Proceedings of Sinn und Bedeutung* 17. 509–526.

Schwarz, Norbert & Fritz Strack. 1990. Context effects in attitude surveys: Applying cognitive theory to social research. In Wolfgang Stroebe & Miles Hewstone (eds.), *European Review of Social Psychology* 2. 31–50. Chichester: Wiley. DOI: https://doi.org/10.1080/14792779143000015

Sharvy, Richard. 1980. A more general theory of definite descriptions. *The Philosophical Review* 89(4). 607–624. DOI: https://doi.org/10.2307/2184738

Snape, Neal & Hironobu Hosoi. 2018. Acquisition of scalar implicatures: Evidence from adult Japanese L2 learners of English. *Linguistic Approaches to Bilingualism* 8(2). 163–192. DOI: https://doi.org/10.1075/lab.18010.sna

Sullivan, Gail M. & Richard Feinn. 2012. Using effect size—or why the p value is not enough. *Journal of Graduate Medical Education* 4(3). 279–282. DOI: https://doi.org/10.4300/JGME-D-12-00156.1

Syrett, Kristen, Anne Lingwall, Silvia Perez-Cortes, Jennifer Austin, Liliana Sánchez, Hannah Baker, Christina Germak & Anthony Arias-Amaya. 2017b. How Spanish-English bilingual children approach entailment-based scalar implicatures. *Glossa Special Issue: Acquisition of Quantification* 2(1). 31. 1–19. DOI: https://doi.org/10.5334/gjgl.76

Syrett, Kristen, Jennifer Austin, Liliana Sánchez, Christina Germak, Anne Lingwall, Silvia Perez-Cortes, Anthony Arias-Amaya & Hannah Baker. 2017a. The influence of conversational context and the developing lexicon on the calculation of scalar implicatures: Insights from Spanish-English bilingual children. *Linguistic Approaches to Bilingualism* 7(2). 230–264. DOI: https://doi.org/10.1075/lab.14019.syr

Tomlinson, John M., Jr., Nicole Gotzner & Lewis Bott. 2017. Intonation and pragmatic enrichment: How intonation constrains ad hoc scalar inferences. *Language and Speech* 60(2). 200–223. DOI: https://doi.org/10.1177/0023830917716101

Torchiano, Marco. 2016. Effsize: Efficient Effect Size Computation. https://cran.r-project.org/web/packages/effsize/index.html.

Uegaki, Wataru. 2015. *Interpreting questions under attitudes*. MIT, MA dissertation.

Vacha-Haase, Tammi. 2001. Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. *Educational and Psychological Measurement* 61. 219–224. DOI: https://doi.org/10.1177/00131640121971194 van Herten, Marieke, Herman H. J. Kolk & Dorothee J. Chwilla. 2005. An ERP study of P600 effects elicited by semantic anomalies. *Cognitive Brain Research* 22(2). 241–255. DOI: https://doi.org/10.1016/j.cogbrainres.2004.09.002

van Rooij, Robert. 2003. Questioning to resolve decision problems. *Linguistics and Philosophy* 26(6). 727–763. DOI: https://doi.org/10.1023/B:LING.0000004548.98658.8f

Vasishth, Shravan & Andrew Gelman. 2017. The illusion of power: The statistical significance filter leads to overconfident expectations of replicability. In *Proceedings of Cognitive Science Conference*. London. https://arxiv.org/abs/1702.00556.

Wilson, Elspeth Amabel. 2017. *Children's development of quantity, relevance and manner implicature understanding and the role of the speaker's epistemic state*. Dissertation. DOI: https://doi.org/10.17863/CAM.17152

Xiang, Yimei & Alexandre Cremers. 2017. Mentions-some readings of plural-marked questions: Experimental evidence. In Andrew Lamont & Katerina A. Tetzloff (eds.), *Proceedings of North East Linguistics Society 47* 3. 261–274. University of Massachusetts, Amherst, MA.

Zhao, Ming, Tao Liu, Gang Chen & Feiyan Chen. 2015. Are scalar implicatures automatically processed and different for each individual? A mismatch negativity (MMN) study. *Brain Research* 1599. 137–149. DOI: https://doi.org/10.1016/j.brainres.2014.11.049

Zhou, Peng, Stephen Crain & Likan Zhan. 2012. Sometimes children are as good as adults: The pragmatic use of prosody in children's on-line sentence processing. *Journal of Memory and Language* 67(1). 149–164. DOI: https://doi.org/10.1016/j.jml.2012.03.005

Zimmermann, Malte. 2007a. Wer ist so alles zum Vortrag gekommen? – Eine semantische Analyse der quantifizierenden w-Fragepartikeln *so* und *alles* im Deutschen. *Handout for the Talk at the GK Satzarten*. University of Frankfurt.

Zimmermann, Malte. 2007b. Quantifying question particles in German: Syntactic effects on interpretation. In Estella Puig-Waldmüller (ed.), *Proceedings of Sinn und Bedeutung (SuB)* 11. 627–641. Barcelona, Universitat Pompeu Fabra.

Zimmermann, Malte. 2010. Quantifying question particles and the non-exhaustiveness of wh-questions. Ms. Potsdam University.

How to cite this article: Fekete, István, Petra Schulz and Esther Ruigendijk. 2018. Exhaustivity in single bare *wh*-questions: A differential-analysis of exhaustivity. *Glossa: a journal of general linguistics* 3(1): 96. 1–32, DOI: https://doi.org/10.5334/gjgl.549

Submitted: 19 October 2017 **Accepted:** 11 June 2018 **Published:** 06 September 2018

]u[*Glossa: a journal of general linguistics* is a peer-reviewed open access journal published by Ubiquity Press. **OPEN ACCESS**