Does ‘a couple’ pattern with scalars or numbers - Insights from inference and ‘so’ tasks

Erying Qin, Chao Sun & Richard Breheny*

Abstract. Previous research establishes that paucal quantifiers like ‘a couple’ are ambiguous between the literal meaning of ‘at least two’ and the enriched meaning understood as conveying a restriction on quantity, the latter of which can be explained by a pragmatic phenomenon, i.e. scalar inference (SI). To address whether this ambiguity patterns with that of scalars or numbers, our Experiment 1 explored the behaviours of ‘a couple’ and scalars with two types of probe questions in inference tasks, and Experiment 2 continued this theme by testing the naturalness rating for ‘a couple’ and scalars in an ‘X so not Y’ construction. The results of our experiments indicate two natures of ‘a couple’: a non-monotonic /cardinal (approximately two) and proportional (a small proportion of).

Keywords. paucal quantifier; scalar inference

1. Introduction. Often hearers go beyond the literal meaning of what speakers utter and make inferences to enrich the message during language comprehension. Scalar inferences (SIs) are a widely discussed example of this kind of phenomenon:

(1) a. Some of the players scored.
b. All of the players scored.
c. Not all of the players scored.

Regardless of whether thought of as grammatical or pragmatic, it is assumed that scalar inference is an operation which can augment the meaning of an utterance beyond what can be derived from its literal meaning. The operation involves the exclusion of a licenced alternative proposition (Horn 1972, Fox & Katzir 2011). In the case of (1a), we assume that the literal meaning of the noun phrase ‘some of the players’ is a monotone increasing, existential quantifier function. Then the sentence in (1b) can be an alternative and its exclusion leads to the implication in (1c). Thus, the sentence in (2a) can optionally be understood as implying (2c).

Going beyond the classic <some, all> example, much recent research has looked at a wider range of cases which can be given a similar analysis (van Tiel et al. 2016, van Tiel & Schaeken 2017). When it comes to noun phrases (NPs) which contain numerals, we can detect a similar duality of readings. Consider (2):

(2) a. Spain has scored two goals
b. If Spain has scored two goals, they have turned the tide of the match.
c. According to the scoreboard, Spain has scored two goals.

The sentence in (2a), when embedded in different contexts in (2b) and (2c), can be understood in different ways. In (2b) we could gloss the understanding as a lower-bounding, ‘two or more goals’, while in (2c), the gloss would be an ‘exactly’ reading – ‘two goals and no more’. If we extend the standard account of ‘some’ to numerals, the numeral ‘two’, when understood to have a literal meaning in the ‘two or more’ sense, would account for (2b). Then a suitable alternative would be a sentence with ‘three’ and the result of combining the literal meaning with the alternative’s exclusion is the ‘exactly’ reading, prominent in (2c). This ‘standard’ account of the numeral case has long been challenged due to the sense that NPs with

* Erying Qin, University College London (erying.qin.18@ucl.ac.uk) Chao Sun, Peking University (chaosun@pku.edu.cn) & Richard Breheny, University College London (r.breheny@ucl.ac.uk).

numerals behave differently to those with ‘some’ and other expressions which are thought to be open to strengthening through scalar inferencing (Horn 1992, Geurts 2006, Breheny 2008). Moreover, a growing body of experimental research , e.g. Marty et al. (2014), Sun & Breheny (2022), points to differences in outcomes for tasks when ‘some’ and numerals like ‘two’ are compared and these results are in line with the view that NPs with numerals do not receive an ‘exactly’ reading as a result of strengthening by scalar inferencing. In this paper, we ask how the paucal quantifier involving ‘a couple’ behaves in terms of the likelihood of SI. For example, ‘a couple’ can mean more or less the same thing as ‘two’, but it is also often used for a broader range of cardinal values than just two, depending on certain properties the objects in question are perceived to have. Consider (3):

(3) a. She scored (only) a couple of goals
b. (Only) a couple of fans in the crowd cheered.

In (3a), the suggestion is that the quantity of goals scored was just two. In (3b), the quantity could be more, with the main implication being that the number is low relative to some standard. Regarding the latter reading Marty & Nevins (under review) report studies showing that, in relatively neutral contexts, participants are prepared to judge quantities far higher than two as counting as ‘a couple’, as long the proportion is low relative to the whole. Marty & Nevins report similar results when the explicit operator ‘only’ is used, suggesting that a willingness to accept ‘a couple’ with larger numbers is not a result of a monotonic, ‘at least a couple’ meaning. One way to capture these data, supported by Marty & Nevins, is to assume that the quantity denoted by ‘a couple’ is only fixed relative to a context (to be a relatively small number). On this analysis, NPs containing ‘a couple’, like ‘some’ are best analysed as existential, upward monotone quantifiers. However, one still needs to account for the fact that many participants judge sentences like in (3a,b) as false when the intersection of restrictor and scope has a cardinality larger than the (contextually determined) small number – as shown in Marty & Nevins. This suggests an accessible upper bounded meaning (‘a couple but not many’). Marty & Nevins assume this upper-bounded meaning arises through scalar inferencing with the Alternative being ‘many’. An alternative would be to align ‘a couple’ with numerals and argue that the two kinds of interpretation of ‘a couple’ NPs arise via different mechanisms.

In our investigation of paucal quantifiers, we capitalize on various experimental paradigms to address the question whether ‘a couple’ patterns with ‘some’ or numerals. In the following part of the paper, we present two experiments: Experiment 1 is based on Sun & Breheny’s (2022) inference tasks (Section 2); Experiment 2 is based on Sun et al.’s (2018) ‘so’ task, which measures the naturalness of an SI-enriched meaning under negation (Section 3).

2. Experiment 1. Sun & Breheny (2022) tested how the quantifier scale <some, all>, the modal scale <possible, certain>, and the numerical scale are interpreted, and established that genuine scalars ‘some’ and ‘possible’ are sensitive to a manipulation that can change the contextual relevance of alternatives (‘all’ and ‘certain’), whilst ‘exactly’ readings of numbers are not. Sun & Breheny (2022) investigated numerals, ‘possible’ and ‘some’ in inference tasks with two types of probe questions. One type, referred to as ‘not Alt’ probe, was intrinsically a standard inference task where the probe question asked participants whether they could infer the negation of a scalar alternative (e.g. not all), according to a speaker character’s statement containing a scalar expression (e.g. some), and Target Response corresponding to inferring the SI was a ‘Yes’ response. The other type of probe question, called ‘could Alt’ probe, asked participants whether, for instance, ‘all’ might not be excluded for the same statement, and Target Response was a ‘No’ response. Note that participants could also give a ‘No’ response when they were uncertain about the speaker character’s intended meaning, irrespective of the probe type. In light of Figure 1 (Sun & Breheny, 2022; p. 9, Figure 4), the interpretations of

'some' and 'possible' were affected by the manipulation of probes, because there were more Target Responses for 'not Alt' than 'could Alt' probes.

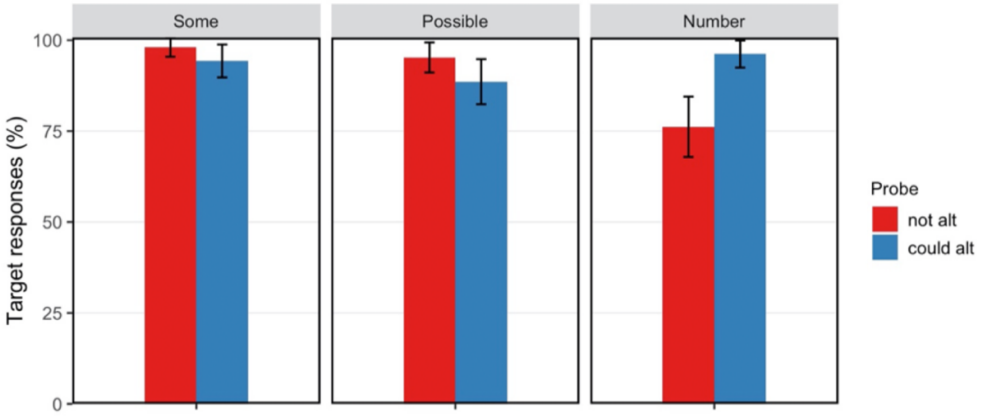

Figure 1: Percentages of target responses for each scale and probe type (Sun & Breheny, 2022, p.9)

This suggested that probe questions had an effect on making the SI contextually relevant, so participants were more certain about inferring the SI as part of the intended meaning, which led to more Target Responses for the 'not Alt' probe. Responses to numbers in their study showed a reversed pattern, indicating that making the contextual question more salient had no effect on target rates. Sun & Breheny explain the reverse pattern of results for numerals as resulting from the fact that the perceived ambiguity for numerals simply leads to an across the board increase in back-off 'no' responses due to uncertainty. This leads to more target responses for 'could Alt' and fewer for 'not Alt' – the attested pattern. Our Experiment 1 mirrors this study so as to see which effect manipulating contexts has on interpreting 'a couple'.

2.1. PARTICIPANTS. 60 native speakers of English participated in an online experiment run on Gorilla Experiment Builder (Anwyl-Irvine et al. 2018). Participants were recruited from Prolific Academic and compensated £0.8. All of them were naïve to the purpose of the experiment. Participants were provided with an electronic version of informed consent before taking part, and this experiment was approved by the UCL research ethics committee.

2.2. MATERIALS AND PROCEDURE. This experiment was a 2×3×3 inference task (probe type × condition × scale), and we manipulated condition and scale within subjects but probe type between subjects (we will elaborate these three factors below).

To avoid the possibility that interpretations of 'a couple' are influenced by characteristics of numerals, we used the experimental items with <a couple, many> to substitute those with numerals in Sun & Breheny's (2022) original study (see Figure 2).

Mary says: *A couple of chairs are in the room.* Would you conclude from this that, according to Mary, *not many chairs are in the room?* ☐ Yes ☐ No	Mary says: *A couple of chairs are in the room.* Would you conclude that, it could be that Mary thinks, *many chairs are in the room?* ☐ Yes ☐ No

Figure 2: Examples of 'not Alt' (left) and 'could Alt' (right) probes

The current experiment tested three scales: <some, all>, <possible, certain> and <a couple, many>. The reason why we chose 'many' as the Alternative is that when accounting for the two readings of 'a couple' in terms of SI-based approach, Marty & Nevins (under review)

regarded ‘many’ as its scalar alternative. For each scale, we constructed target, ‘true’ control and ‘false’ control conditions, and there were 6 items for target condition and 12 items for control condition (6 ‘true’ control items and 6 ‘false’ control items). In short, control items had the same structure as target items, except for the conclusions, the responses of which in the control conditions were either clearly ‘Yes’ or clearly ‘No’.

As probe type was a between-subject factor, each participant was randomly assigned to one of the two probe types and saw 54 items, including 6 target items, 6 ‘true’ control items and 6 ‘false’ control items per scale. Two lists, the list of ‘not Alt’ probe and that of ‘could Alt’ probe, were created. Each item only appeared once in each list, and the order of items was randomised for each participant in each trial. The inference tasks started with instructions and four practice trials.

2.3. RESULTS AND DISCUSSION. Participants were removed if their accuracy on control items was below 70%. Three participants in the ‘not Alt’ group and nine participants in the ‘could Alt’ group were removed, and the overall mean accuracy of the control items reached 93% (‘true’ control condition: 89%, ‘false’ control condition: 97%).

Putting aside the control condition, we coded the ‘Yes’ response to the ‘not Alt’ probe and the ‘No’ response to the ‘could Alt’ probe as Target Response. Figure 3 shows the percentages of Target Responses for each scale and probe type. To analyse these Target Responses, we constructed a mixed effects logistic regression model predicting responses (target vs. non-target) on the basis of probe type (not Alt vs. could Alt), scale type (‘some’ vs. ‘possible’ vs. ‘a couple’), and their interaction, including random intercepts for participants. Random slopes were dropped due to non-convergence or singularity. The mixed-effect analyses, as well as all of the following analyses that will be reported, were conducted in R (R Core Team 2022) using the ‘lme4’ package (Bates et al. 2015). Degrees of freedom and corresponding p-values were estimated using the Satterthwaite’s method, as implemented in the ‘lmerTest’ package (Kuznetsova et al. 2017). Scale was dummy-coded, with ‘some’ as the reference level, and probe was deviation coded. Model comparisons were conducted to test the significance of fixed effects with more than two levels, using likelihood ratio tests. Significant interactions were followed up by conducting analyses on subsets of data defined by the levels of relevant factors.

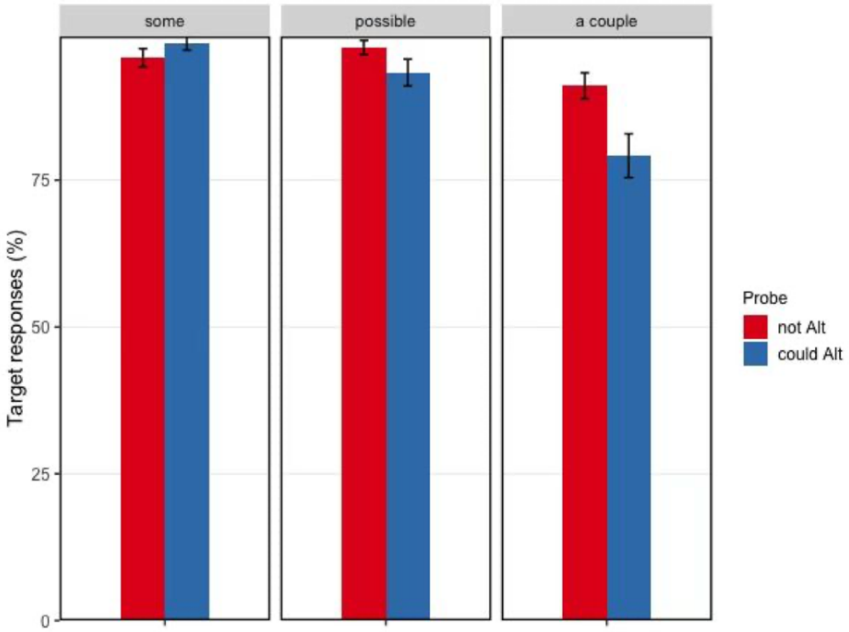

Figure 3: Percentages of target responses for each scale and probe type. Error bars represent standard errors

The interaction between scale and probe type was significant ($p < .001$). On the target responses, there was a main effect of scale ($p = .03$) and a main effect of probe type ($p < .001$), indicating that the differences in the probability of target responses among three scales were greater for 'not Alt' than 'could Alt' probes. Further analyses revealed that for 'a couple', the probability of target responses was higher for the 'not Alt' probe compared to the 'could Alt' probe ($p < .01$). The same effect was only marginally present for 'possible' ($p = .07$), but there were no statistically significant differences between the 'not Alt' and the 'could Alt' probes for 'some' ($p = .23$). Our results were consistent with the pattern in Sun & Breheny between 'not Alt' and 'could Alt' probes for 'possible'. In terms of 'a couple', the significantly different probability of Target Responses between the 'not Alt' probe and the 'could Alt' probe suggested that paucal quantifiers, such as 'a couple', behave like a genuine scalar expression, not like numbers.

If we look at the data more closely, however, for 'a couple', the rate of Target Responses to 'not Alt' probe was significantly lower than that for 'possible' ($p = .04$), which was similar to that found by Sun & Breheny when numerals were compared to scalars in 'not Alt' trials . For the 'could Alt' probe, the rate of target responses was lower for 'a couple' than for the other two scalars (some: $p < .001$; possible: $p < .001$) and there was no statistically significant difference between 'possible' and 'some' ($p = 0.4$). Sun & Breheny argue that the lower rate in 'not Alt' trials for numerals than scalars is indicative of the fact that at least some participants recognized the ambiguity of the target trial sentence and became more non-committal. A similar drop off in rates for 'a couple' compared to the other scalars might indicate a kind of ambiguity between a non-monotonic, 'exactly/approximately two' interpretation and a proportional, 'a small number of' interpretation. The latter interpretation, like 'some' and 'a few', may be open to scalar inference, while the former, like numerals, not so.

3. Experiment 2: naturalness rating for scalar expressions under negation. One distinguishing feature of numerals compared to many other widely discussed scalar expressions lies in their behaviour in linguistic contexts which tend to block scalar inference, such as in the scope of negation (Horn 1992, Breheny 2008). To illustrate, while (4a) is readily accepted in a case where she ate more than two cookies (say, three), (4b,c) are not readily acceptable where the stronger term is true; i.e. where it is certain she ate cookies in the case of (4b) or where she ate both a cookie and a cake in the case of (4c):

(4) a. She did not eat two cookies.
b. It is not possible she ate cookies.
c. She didn't eat a cookie or a cake.

Sun et al. (2018) report a study from which they extract a measure of felicity of scalar inference strengthening under negation. As per (4a) above, numeral expressions are quite felicitously understood in a non-monotonic sense in the scope of negation, while 'possible' and 'or' are not. Sun et al. devised a 'S so not W' probe in order to collect felicity judgements for this. The idea behind the probe is that the sentence would be infelicitous unless there is local strengthening under negation. The 'S' term is simply an alternative that unilaterally entails the putative literal meaning of a monotone scalar expression. For example, for 'possible' we can use 'certain', as in (5b) below:

(5) a. She ate three cookies, so not two.
b. It is certain that she ate cookies, so not possible.
c. The weather is hot, so not warm.

Assuming that scalar strengthening is blocked, or strongly disfavoured, under negation, the resulting 'S so not W' sentence should appear incoherent. But, to the extent that the non-monotonic meaning is permissible under negation, the sentence should strike participants as

less infelicitous (see Breheny 2008 for more discussion). Intuitively, this seems to be the case with (5a) above. Sun et al. employed 42 different scalar terms, including 'possible', 'warm' and all others taken from van Tiel at al. (2016). Sun et al. took graded responses to 'S so not W' sentences as a measure of 'liability to strengthen under negation' and found that this measure can account for previously unexplained variance in rates of target responses in their replication of van Tiel et al's inference task study.

Sun et al.'s stimuli did not include numerals or 'a couple', and the highest rated scalar expressions on their 'S, so not W' task were only at the mid-range of a seven-point Likert scale (see Figure 4 below). In our second experiment, we wanted to bring numerals into the stimulus set and to explore more the idea that 'a couple' might have two kinds of interpretation, one that is more like 'some' and other scalar expressions, and one that is more like numerals. In order to do this, we re-considered what might be the alternative expression to use for 'a couple' in the stimuli. Recall that in Experiment 1, we used 'many' as the alternative for 'a couple', following on from Marty & Nevins. Intuitively, the felicity of 'many' as an alternative for 'a couple' relies on the quantifier being understood in its proportional sense (see Partee 1989). As discussed in relation to Experiment 1 above, one way to account for the results would be to suppose that there is a second sense for 'a couple' which is more like 'exactly/approximately two'. When thinking about the 'S, so not W' probe for 'a couple' we assumed that if the 'S' term were a partitive form involving 'many' then that would prime the proportional, scalar meaning, while if a non-partitive numerical NP played the role of 'S', that would better prime any small-number approximative sense. These ideas were implemented in the design below. If we are right about participants having these two ways to understand 'a couple', we expected to see different outcomes using the different expressions in the 'S' role. Specifically, when 'S' is a numeral, then felicity of 'a couple' should be more like that of numerals, compared to when 'many' is used.

3.1. PARTICIPANTS. 103 native speakers participated for £1.5 compensation. Recruitment and screening were identical to Experiment 1.

3.2. MATERIALS AND PROCEDURE. We used 48 scalars including 43 of them investigated in Sun et al. 's (2018) study along with numerals, 'a couple/ number', 'a couple/many' and some other scalars to construct experimental sentences for Experiment 2. The experimental sentences were of the form 'X so not Y,' where X and Y were chosen according to the principle outlined above. Figure 4 is an example item for 'some':

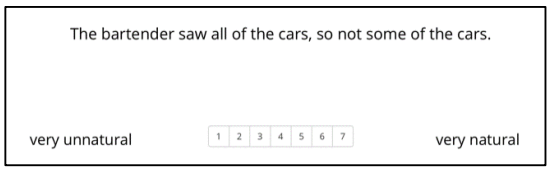

Figure 4: Example displays for 'some'

In the case of 'a couple' the two kinds of item are illustrated in Figure 5 below:

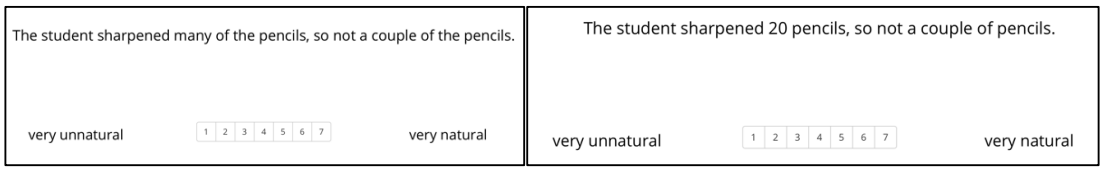

Figure 5: Examples of partitive (left) and non-partitive (right) groups

We employed a between-subject design involving a partitive group and a non-partitive group. All the scalars in the two groups were the same, except for 'a couple/number' in the

non-partitive group, whilst ‘a couple/many’ in the partitive group. Each participant was randomly assigned to one of the two groups and judged 47 experimental sentences. Participants were asked to indicate how natural these constructions are on a 1 (very unnatural) - 7 (very natural) Likert scale. In addition to experimental sentences, participants also had a chance to use the extremes of the Likert scale for felicitous fillers (e.g. The window is open so not closed.) and infelicitous fillers (e.g. The train arrived so it never departed.).

3.3. RESULTS AND DISCUSSION. One participant in the non-partitive group was excluded because the mean ratings for the infelicitous/filler items were above 5. Again, the data analysis was performed using R (R Core Team 2022). The overall results for the two groups here compared with scales in Sun et al. (2018) are shown in Figure 6. When comparing our outcomes to those in the original study, we observed no significant difference in the mean ranks (Wilcoxon signed-rank test: $p = 0.089$), thus the ‘so’-task results broadly align with those in Sun et al. (2018). We note that overall, ratings for the 43 items that were common between Sun et al’s previous study and this study were lower here. We assume this is due to the fact that participants tend to fix their range of ratings in relation to items that they have already seen and in this new experiment, the two new items involving numerals and ‘a couple’ were generally much more felicitous, making the other items seem less felicitous as a result.

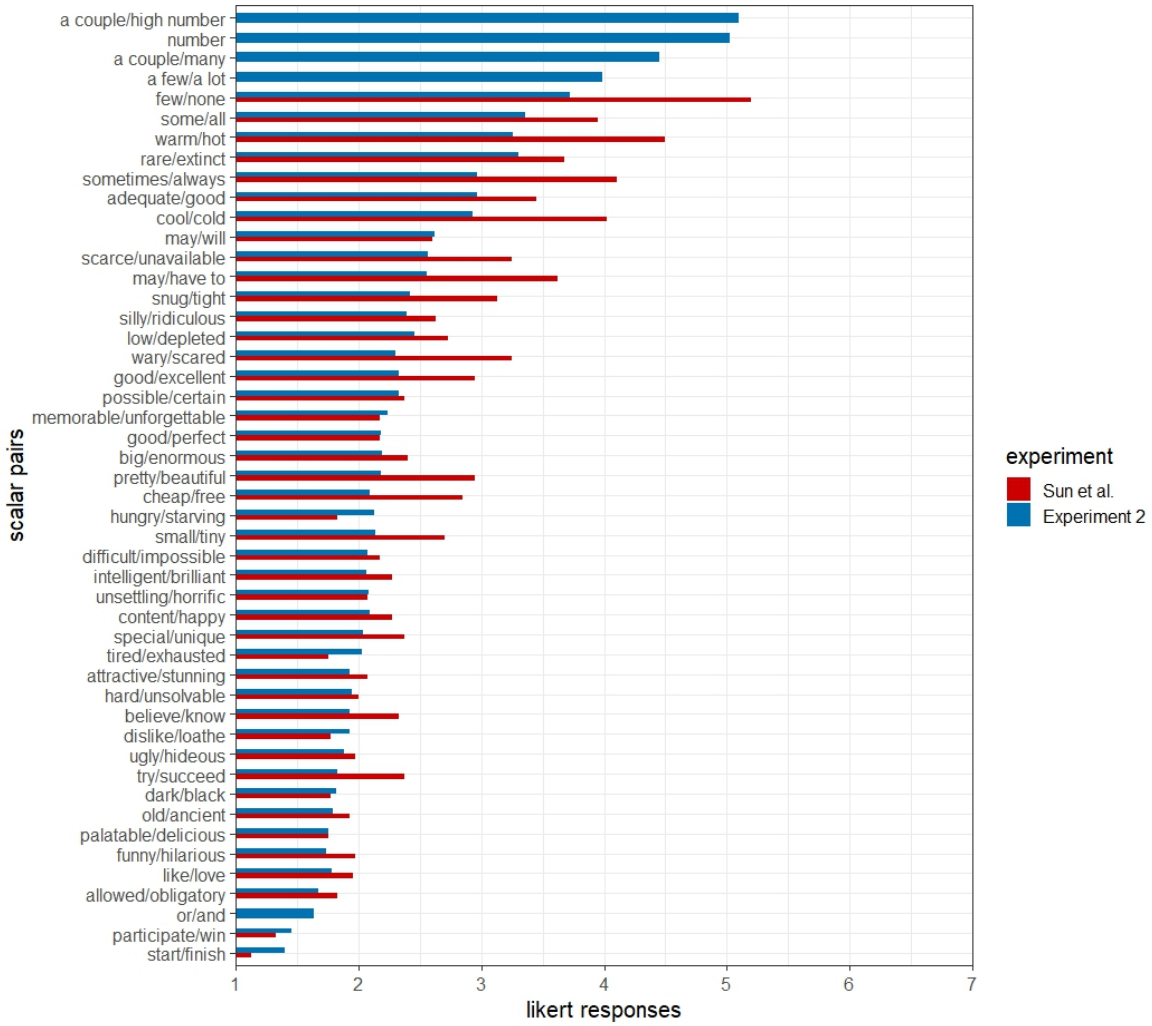

Figure 6: Mean naturalness ratings for ‘so’ task. Note that ratings shown for all scales except for ‘a couple’ is the average across the two groups.

Turning now to the scalar expressions of interest, Figure 7 shows the ratings from each group for stimuli with numerals, 'a couple' and three quantifier expressions which were highly ranked in both groups (and in Sun et al's previous study). We can see that, as expected, 'so'-task probes with numerals had a high felicity rating – higher than 'a few', 'few' and 'some'. The question of interest for us is the comparative felicity of the probes for numerals and 'a couple' between the two groups.

To compare the behaviour of 'a couple' and numerals in the two conditions, the Mann Whitney U test was conducted. We find a significant difference, when comparing 'a couple/many' to number ($p = .05$). However, there was no significant difference between 'a couple/high number' and number ($p = 0.7$). Thus, the different expressions playing the 'S' role in the probe for 'a couple' had the expected effect. Participants who saw a number like '20', their judgement about the felicity of the probe was not different to the probe for numeral. When participants saw a sentence like 'many..., so not a couple' they found the sentence significantly less felicitous than the numeral items.

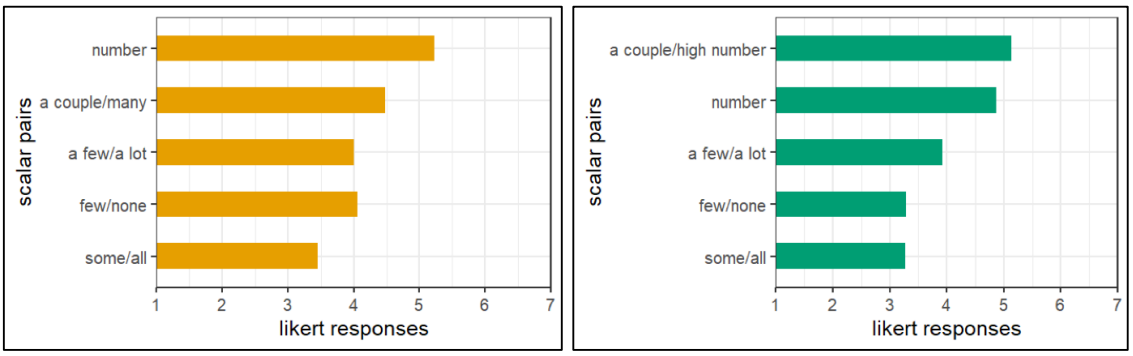

Figure 7: Mean naturalness ratings for 'so' task in the partitive (left) and the non-partitive (right) group

4. Summary. Experiment 1 provides mixed evidence that 'a couple' patterns with genuine scalar items such as 'some'; however, lower rate of target response to 'a couple' in 'not Alt' condition compared to other scalars was similar to that of numerals in previous research. Based on these results, we speculated that participants may have more than one way to interpret noun phrases with 'a couple'. One way is more like those with 'some' and other existential monotone increasing quantifiers. One way is more like numerals, which have widely been viewed as not behaving in this way.

Turning now to Experiment 2, we wanted to exploit the known felicity of the non-monotonic interpretation of numerals in the scope of negation as a means to explore the possibility of two readings for 'a couple'. Using the 'so'-task developed in Sun et al. (2018), we devised two probe stimuli for 'a couple' each of which we expected to emphasise a different one of the two proposed ways that these noun phrases may be understood. The results show that when primed by a numeral alternative, 'a couple' behaved essentially in the same way as numerals in Sun et al.'s 'so'-task. When primed with 'many', participants found the probes less felicitous. Overall, our results are suggestive that 'a couple' may semantically have two aspects. We acknowledge that this interpretation of the results is somewhat indirect and other interpretations are possible. We also leave open here more detailed analysis of how to formally analyse each of these two aspects of 'a couple' and to explain how they may be generated.

References

Anwyl-Irvine, AL, Jessica Massonnié, Adam Flitton, Natasha Kirkham & Jo K. Evershed. 2020. Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods* 52. 388–407. https://doi.org/10.3758/s13428-019-01237-x.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1). 1–48. https://doi.org/10.18637/jss.v067.i01.

Breheny, Richard. 2008. A new look at the semantics and pragmatics of numerically quantified noun phrases. *Journal of Semantics* 25(2). 93–139. https://doi.org/10.1093/jos/ffm016.

Fox, Danny & Roni Katzir. 2011. On the characterization of alternatives. *Natural language semantics* 19 (2011). 87–107. https://doi.org/10.1007/s11050-010-9065-3.

Geurts, Bart. 2006. Take 'Five': The meaning and use of a number word. *Non-definiteness and Plurality* 95. 311–329. https://doi.org/10.1075/la.95.16geu.

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

Horn, Laurence R. 1992. The said and the unsaid. *Semantics and linguistic theory*. 163–192.

Kuznetsova, Alexandra, Per B. Brockhoff & Rune H.B. Christensen. 2017. lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software* 82(13). 1–26. https://doi.org/10.18637/jss.v082.i13.

Marty, Paul, and Andrew Nevins. Under review. Expressions of Paucity: Where is the Upper-Bound?. Ms UCL

Partee, Barbara. 1989. Many quantifiers. *Proceedings of the 5th Eastern States Conference on Linguistics* 5. 383–402.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*.

Sun, Chao & Richard Breheny. 2022. The role of Alternatives in the interpretation of scalars and numbers: Insights from the inference task. *Semantics and Pragmatics* 15(8). 1–15. https://doi.org/10.3765/sp.15.8.

Sun, Chao, Ye Tian & Richard Breheny. 2018. A link between local enrichment and scalar diversity. *Frontiers in Psychology* 9. https://doi.org/10.3389/fpsyg.2018.02092.

van Tiel, Bob, Emiel van Miltenburg, Natalia Zevakhina & Bart Geurts. 2016. Scalar diversity. Journal of semantics 33(1). 137–175. https://doi.org/10.1093/jos/ffu017.

van Tiel, Bob & Walter Schaeken. 2017. Processing conversational implicatures: alternatives and counterfactual reasoning. *Cognitive science* 41. 1119–1154. https://doi.org/10.1111/cogs.12362.