Preferences in the quantified description of visual groups

Gordon Briggs[1], **Hillary Harner**[1,2], and **Sangeet Khemlani**[1]
{gordon.briggs, hillary.harner.ctr, sangeet.khemlani}@nrl.navy.mil
[1] Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory, Washington, DC 20375
[2] NRC Postdoctoral Fellow

Abstract

Research suggests that people minimize the amount of effort used to generate natural language descriptions of visual scenes. In the case of visual scenes with multiple groups, recent work has found that people tend to generate quantitative descriptions that mention the number and cardinality of groups (e.g., "two groups of three limes"), but omit the total quantity (e.g., "six limes"). This finding suggests that people groupitize, that is, they more quickly determine the number of grouped items by rapid enumeration of subgroups, rather than slower item-by-item counting. A recent proposal predicts that during description, people exert less effort by encoding and reporting only information that is readily available to perception. In previously studied description tasks, people may have omitted the total quantity from their descriptions because of considerations of brevity and informativity. In this paper, we describe a study designed to test how individuals balance effort, brevity, and informativity when evaluating quantified descriptions. The experiment was designed to elicit more fine-grained preferences for descriptions using direct comparisons between two competing descriptive forms. The results suggest that perceptual effort plays a central role in how people describe grouped collections of items.

Keywords: numerical perception; pragmatics; quantified description; visual grouping

Introduction

When communicating about visual scenes with multiple objects in them, people often produce quantified descriptions. Consider the scene depicted in Figure 1. Many possible options exist to accurately describe the image, e.g., "six limes", "several limes", "more than four limes". The relevant goals of discourse, and other pragmatic constraints, serve as guides for generating descriptions (Cummins, 2015; Hesse & Benz, 2018).

During production of quantified descriptions, pragmatic constraints are often in tension with perceptual constraints. For example, when placed under time constraints that make exact enumeration difficult, people produce inexact quantified descriptions of sets of items (Briggs, Wasylyshyn, & Bello, 2019). In a reference generation task, where people use quantified expressions to distinguish one set of objects from other sets of different quantities, a similar modulation of precision occurs (Barr, van Deemter, & Fernández, 2013). Specifically, Barr et al. (2013) found that people use exact numbers to describe items below the subitizing range, which is the range of quantities that people can exactly enumerate without counting. Many researchers suggest that the subitizing range is from one to around four items, meaning that people can rapidly and accurately enumerate quantities of 4 or fewer (Mandler & Shebo, 1982; Trick & Pylyshyn, 1994). Barr et al. (2013) also found that people use non-numerical quantifiers when describing quantities outside the subitizing range. These findings are consistent with the *least perceptual effort hypothesis* (Briggs, Harner, & Khemlani, 2020), which posits that speakers exert the minimum amount of perceptual effort to accomplish a particular task. Thus, for small subitizable quantities, where exact enumeration is low-effort and automatic, people should produce exact number descriptions. For large quantities, when exact enumeration is high-effort, people should produce less precise quantified descriptions. In other words, the hypothesis predicts that pragmatic pressures to be precise and informative interact with the level of perceptual effort the speaker is willing to exert.

Figure 1: An example of a visual scene with multiple groups of similar objects.

The phenomenon of quantified description becomes more complex when a scene has multiple groups. Consider again the image in Figure 1. People can describe not only the total number of limes, but also the number of groups of limes (i.e., two) and the cardinality of each of these groups (i.e., three). Inclusion or exclusion of these three different quantified referents yield several possible forms of quantified descriptions. Briggs et al. (2020) found that for certain scenes that depict multiple groups of items, people tend to generate quantified descriptions that describe the number and cardinality of the groups, but omit the total quantity (e.g., two groups of tree limes). The scenes that yield such descriptions are those in which the number of groups and their cardinalities are at or below the subitizing limit, such as in Figure 1.

Research in numerical perception has shown that "groupitizing" is a common means to determine the exact quantity of a set of objects that can be easily decomposed into subgroups of small quantities (Starkey & McCandliss, 2014; Ciccione & Dehaene, 2020; Anobile, Castaldi, Maldonado, Burr, & Arrighi, 2020). People can, for example, determine the number of limes in Figure 1 by first noticing that there are two groups of limes, and that both have three limes. Simple mental arithmetic allows the viewer to determine the total quantity without relying on item-by-item counting. Thus, according to the least perceptual effort hypothesis, people should produce quantified descriptions that include only the group number and the group cardinality. These descriptions may serve pragmatic considerations as well. For example, saying "two groups of three limes" is briefer than saying "six limes in two groups of three each," and more informative than saying "six limes."

In this paper, we examine the role of perceptual and pragmatic factors in the quantified description of visual scenes with multiple groups. We begin by introducing three key perceptual and pragmatic constraints and discuss what predictions they make regarding what quantified description forms should be favored over others. In particular, we identify a particular pairwise comparison that makes a prediction unique to the least perceptual effort hypothesis. We then present a novel experiment designed to elicit description preferences given direct comparison between two competing descriptive forms, and show that participants' preferences corroborate the prediction. We conclude by discussing the results of the experiment and how perceptual and pragmatic constraints interact in visual scene description.

Preferences in descriptions of groups

Perceptual and pragmatic constraints can account for preferences between forms of quantified descriptions of visual groups. At least three kinds of quantification are relevant when considering quantified descriptions of groups: the total number of similar visual items, the number of groups, and the cardinality of each group. Consider the following quantified descriptions of Figure 1:

"There are six limes." [**D1**]

"There are six limes in two groups." [**D2**]

"There are two groups of three limes." [**D3**]

"There are six limes in groups of three." [**D4**]

"There are two groups of three limes for a total of six limes." [**D5**]

Perceptual and pragmatic constraints make diverging predictions as to which types of quantified description people should favor over others during natural language production, so we review their differences.

Perceptual constraints. When the number of groups in an image is easy to enumerate, i.e., below the subitizing limit, people appear to have rapid access to that number. And when the number of items in a particular group is also below the subitizing range, people should have access to the group cardinality as well. Hence, if people base descriptions on whatever numbers come to mind most rapidly as a result of perceptual processes, as the least effort hypothesis predicts, they should favor descriptions such as D3 over more complex descriptions (such as D4 and D5). They should also favor D3 over simpler descriptions such as D1, because D1 is based on the number of total items, i.e., a number that may not be rapidly available.

Pragmatic constraints. People may favor expressions that are more informative or briefer than alternative expressions. These constraints correspond to the Gricean maxim of quantity and manner, respectively (Grice, 1975). We do not consider the constraint of correctness (the Gricean maxim of quality) as the vast majority of people tend to describe images such as Figure 1 accurately. Hence, D1-D5 above are all accurate descriptions, and people cannot evaluate them based on their relative accuracies.

Considerations of informativity concern those pieces of quantitative information that can be communicated or easily inferred. Hence, if viewers are concerned with informativity alone, they should prefer descriptions D3, D4, and D5 as the most informative. D5 is most informative because it explicitly communicates all 3 pieces of quantitative information, whereas with D3 and D4, the quantitative information not directly communicated can be inferred from the other two descriptors. In the case of D3, the total quantity can be inferred, while D4 provides enough information to infer the number of groups. D1 is the least informative form, because while it conveys the total number of items, it does not convey any indication that the items are grouped (let alone the number of groups or the cardinality of those groups). D2 is more informative than D1, as it mentions the number of groups, but it does not convey any information about the cardinality of each group. For instance, a description such as "six limes in two groups" could describe an image where one group has two limes and the other four, not distinguishing it from an image where both groups have three limes.

Considerations of brevity, i.e., a constraint that pressures viewers to produce concise descriptions, apply to the task of quantified descriptions by determining the number of facts explicitly communicated. Brevity could be construed in terms of precise word counts, but various surface realizations of the descriptive forms above could yield expressions of different lengths (e.g., "six limes in two groups" vs. "six limes in groups of two items") and so we construe brevity as a more conceptual constraint. The constraint yields the following ordering: D1 is the briefest description, conveying only one quantitative fact. D5 is the least brief, explicitly conveying all three relevant facts, and D2, D3, and D4 all convey two pieces of quantitative facts. Hence, a constraint on brevity, in

isolation, should favor D1 over D2-D5 and D5 over D2-D4.

Probing preferences. In a prior study, eliciting descriptions of images such as in Figure 2, Briggs and colleagues (2020) found D3 to be the most common form produced by participants, consistent with the least perceptual effort hypothesis. However, this preference is also consistent with pragmatic factors such as informativity (since D3 is more informative than D2 and D1) and brevity (since D3 is more concise that D5). The D3 preference in these cases may be overdetermined. Thus, the elicitation task presented by Briggs and colleagues (2020) cannot unequivocably predict what candidate forms people consider when they generate descriptions. It may be that the task of generating a description biases participants to incorporate considerations of informativity and brevity. Hence, we report a novel experiment that sought to further elaborate on this work by investigating direct comparisons between different forms of quantified description.

Experiment

The aim of the experiment was three-fold. First, it sought to gather more fine-grained data regarding description preferences, including between candidate descriptions that people do not produce often. Second, the study tested preferences between D3 and D4: these two statements are matched on their informativity and brevity, and so the least perceptual effort hypothesis uniquely predicts a preference for D3. Third, it sought to test whether the strong preference for D3 descriptions predicted by the least perceptual effort hypothesis and found in prior work could be replicated in a non-elicitation paradigm.

The experiment investigated descriptive preferences between a limited set of possible candidate expressions. All the images presented multiple groups of objects, with each group containing the same number of objects. Participants viewed two possible descriptions of the image and selected the one they found to be the more natural description (or else indicated that they had no preference between the two descriptions).

Method

Participants. Fifty-three participants (mean age = 36.6 years; 31 males, 21 females, and 1 no response) volunteered through the Amazon Mechanical Turk platform (see Paolacci, Chandler, and Ipeirotis (2010), for a review). All participants reported being native English speakers; we dropped one participant due to a data recording error.

Procedure. Participants carried out 20 trials where they were presented with each of the 10 possible pairwise description comparisons twice. The experiment randomized the order of the trials. On each trial, participants viewed a single image randomly selected from the materials. Below the image, the participants were presented with two descriptions that accurately reflected the quantities contained within the selected image. Each trial was self-paced with no time limit. Participants selected the description they found to be a more natural description of the image, or otherwise selected a third option to indicate 'no preference' between the two descriptions. Figure 2 shows an example trial, where the participant was asked to select between forms D1, D5, and NP (no preference). The presentation order of the two candidate forms was randomized on each trial, while the 'no preference' option was always placed last.

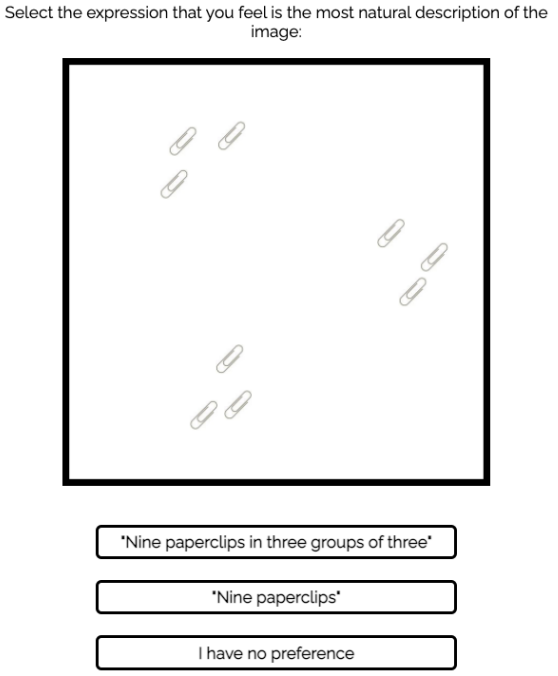

Figure 2: A screenshot from an example trial within the experiment.

Design. The study manipulated the possible descriptive forms (i.e., D1–D5) participants were presented with, and participants acted as their own controls. The study gave participants pairwise choices between two different forms. Therefore, there were $C_2^5 = 10$ possible pairwise comparisons (i.e., {D1,D2}, {D1,D3}, ..., {D4,D5}).

Materials. Images in the study contained multiple homogeneous sets of everyday objects (e.g., paperclips, mugs, flowers). A total of eight different object types were used. Images involved four possible combinations of number of groups and group cardinality, specifically: 2x4, 3x3, 3x4, and 4x3. Figure 3 provides an examples of stimuli from each of these variants. The materials included an image for each object type and group configuration, yielding a total of 32 unique images.

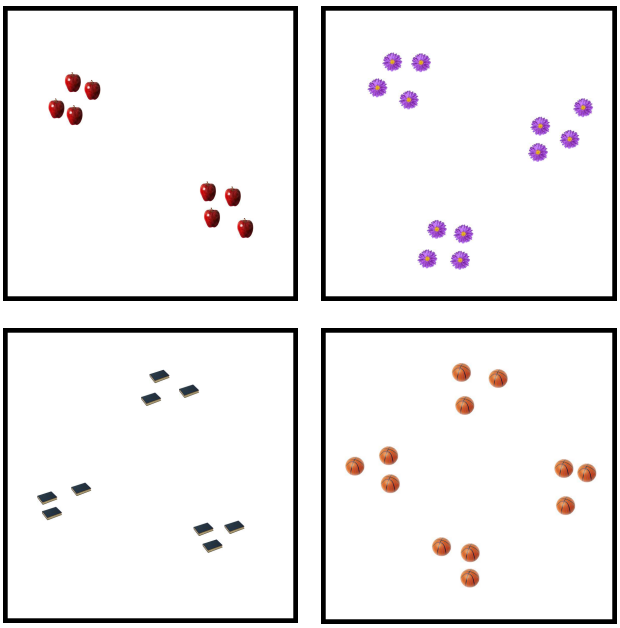

Figure 3: Example stimuli representing each possible group number and group cardinality (i.e., 2x4, 3x3, 3x4, 4x3).

Open science. Data from the experiment, experimental code, and statistical analyses are all available online through the Open Science Framework (`https://osf.io/u7gyc/`).

Results

Table 1 presents the percentages of response preferences for each pairwise comparison of candidate descriptions. To analyze the significance of a response pattern in a given comparison, we coded one response as -1, the other as 1, and the 'no preference' option as 0. The mean responses for each participant were then compared to a null hypothesis sample (all zeros to indicate no preference) using Wilcoxon signed-rank tests. In the analyses reported below, a Bonferroni correction was applied to account for the ten pairwise comparisons involved in the analysis. The corrected significance level was set to $p < .005$.

Least effort hypothesis. Overall, the data supported the least perceptual effort hypothesis. Participants preferred D3 descriptions reliably more often to D4 descriptions (64% vs. 35%; Wilcoxon test, $z = 3.11, p = .002$, Cliff's $\delta = .31$). Likewise, they preferred D3 more often than D1 (70% vs. 28%; Wilcoxon test, $z = 3.65, p < .001$, Cliff's $\delta = .423$). The data were consistent with the least perceptual effort hypothesis for D3 vs. D2 and D3 vs. D5, where D3 was more commonly preferred to D2 (59% vs. 39%; Wilcoxon test, $z = 1.77, p = .077$, Cliff's $\delta = .19$) and D5 (57% vs. 41%; Wilcoxon test, $z = 1.33, p = .18$, Cliff's $\delta = .135$), respectively, though these results were not statistically reliable.

Informativity predictions. The preferences predicted by the informativity constraint were mostly supported by the data. In particular, participants dispreferred D1 compared to all competing descriptions. As previously reported, D1 was significantly dispreferred to D3. Likewise, D1 was significantly less preferred to D5 (29% vs. 67%; Wilcoxon test, $z = 3.08, p = .002$, Cliff's $\delta = .39$) and D2 (29% vs. 67%; Wilcoxon test, $z = 3.15, p = .002$, Cliff's $\delta = .39$). Participants preferred D1 less often than D4, but this comparison was marginal in significance (36% vs. 58%; Wilcoxon test, $z = 1.96, p = .05$, Cliff's $\delta =$). Participants selected D2 more often than D4, contrary to the prediction from informativity. However, this result was not significant (47% vs. 44%; Wilcoxon test, $z = 0.27, p = .79$, Cliff's $\delta = .04$).

Brevity predictions. Participants tended to violate the predictions of the brevity constraint for many comparisons, e.g., they preferred D5 to D4 descriptions (57% vs. 34%; Wilcoxon test, $z = 2.04, p = .041$, Cliff's $\delta = .23$) and D5 to D1 descriptions (see previous section). Those scenarios in which participants preferred briefer descriptions happened to concern the comparisons in which the briefer description was D3, as in D3 vs. D4 or D3 vs. D5. As a whole, the study lends more support to the least effort hypothesis than the brevity constraint.

Summary. Overall, the preferences found in the data were consistent with both the predictions made by the least perceptual effort hypothesis and the informativity constraints. Participants preferred D3 over all other competing descriptive forms, as predicted by the least perceptual effort hypothesis. In particular, participants preferred D3 over D4 responses, a preference uniquely predicted by the least perceptual effort hypothesis. These results reveal that the interaction between perceptual cost and informativity found in other quantified description tasks without multiple groups (Barr et al., 2013; Briggs & Harner, 2019) is robust and generalizable. The data complement the findings from a quantified description elicitation task that involved scenes with multiple groups, which revealed a strong preference for D3 descriptions (Briggs et al., 2020).

General Discussion

When formulating descriptions of visual scenes, speakers must balance a variety of constraints. On the one hand, they ought to be cooperative speakers by satisfying various pragmatic goals: they should be appropriately informative by communicating necessary information and omitting redundant information (Grice, 1975). On the other hand, speakers are constrained by the way they perceive a visual scene, since perception demands time and effort. The least perceptual effort hypothesis proposes that people generate descriptions in a way that minimizes perceptual costs. It predicts that speakers generate descriptions based on the information readily available to them.

In practice, perceiving a scene to know what information is redundant requires additional effort beyond perceiving what is needed to be sufficiently informative. Thus, the

	Percentages of descriptions selected as preferred						
Comparison	D1	D2	**D3**	D4	D5	NP	Significance
D1 vs. D2	28	67	–	–	–	4	*
D1 vs. D3	28	–	**70**	–	–	2	*
D1 vs. D4	36	–	–	58	–	7	.
D1 vs. D5	29	–	–	–	67	4	*
D2 vs. D3	–	39	**59**	–	–	3	
D2 vs. D4	–	47	–	44	–	9	
D2 vs. D5	–	38	–	–	55	8	
D3 vs. D4	–	–	**64**	35	–	1	*
D3 vs. D5	–	–	**57**	–	41	2	
D4 vs. D5	–	–	–	35	58	8	.

Table 1: Percentages of trials on which participants preferred one of the five descriptive forms in Experiment 1 for each pairwise comparison; NP = "no preference". [* = $p \leq .005$; . = $p \leq .05$]

least perceptual effort hypothesis predicts a tendency toward over-informativity. For instance, many researchers have investigated how people produce referring expressions, where a speaker must describe a target item to distinguish it from distractors. Overspecification is a common phenomenon in such cases: speakers describe redundant attributes beyond what is necessary to disambiguate the item (Pechmann, 1989; Tarenskeen, Broersma, & Geurts, 2015; Koolen, Krahmer, & Swerts, 2016; Rubio-Fernández, 2016; Viethen, van Vessem, Goudbeek, & Krahmer, 2017). Likewise, in the case of quantified referring expressions, where target collections are based on quantity information, speakers display a similar tendency toward over-informativity (Barr et al., 2013; Briggs & Harner, 2019), that is, they often describe an image as "4 black dots" when a description such as "a set of dots" is sufficient (Barr et al., 2013).

In this study, we examined preferences in how people evaluate quantified description of visual scenes that contain multiple groups. When multiple groups are present, the space of possible quantified descriptions grows more complex, and it presents viewers with the option of not only reporting the total number of items in a scene, but also the number of groups, and cardinality of each group. Specific pragmatic constraints make predictions about what descriptive forms should be preferred by speakers. For example, the constraint of brevity would favor reporting only the total number of items, whereas informativity would favor the forms that allow hearers to infer the total number of items, the number of groups, and the cardinality of each group. Research in numerical perception has shown that people "groupitize" to determine the exact quantity of a set of objects that can be easily decomposed into subgroups of small quantities (Starkey & McCandliss, 2014; Ciccione & Dehaene, 2020; Wege, Trezise, & Inglis, 2021), i.e., they determine quantity by first encoding the number of groups and cardinality of subgroups, then performing mental computation to obtain the total quantity without relying on item-by-item counting. Thus, according to the least perceptual effort hypothesis, people should commonly produce quantified descriptions that include only the number of groups and their cardinalities.

A previous elicitation study supported this prediction (Briggs et al., 2020). However, a limitation of that study was that it did not provide insight into what candidate forms people considered as they generated descriptions, only the final description produced by the participants. In particular, we sought to advance that work by examining whether or not a description of only group number and cardinality (D3 above) was preferred to a description containing total number and group cardinality (D4 above). This preference is predicted only by the least perceptual effort hypothesis. To obtain such fine-grained data about descriptive preferences, we designed an experiment that asked participants to engage in pairwise comparisons of descriptive forms. The data supported the preference to report group number and cardinality predicted by the least perceptual effort hypothesis.

One of the limitations of the present study is that it only deals with groups of homogenous cardinality and group cardinalities within the subitizing range. The prediction of reporting both the exact number and cardinality of groups may not apply in such situations. For instance, consider an example of a collection of items with subgroups containing 10 items each. In this case, the least perceptual effort hypothesis predicts that people may choose to describe the group cardinality imprecisely (e.g., "groups with lots of dots") or not at all. And, when the cardinality of each subgroup differs, precise descriptions of the cardinality of each group may become both harder to perceive and encode and unwieldy to produce. The pairwise comparison paradigm reported in this study may help investigate preferences that are not predictive of elicited descriptions. For example, it may be the case that participants favor exact descriptions of quantities outside of the subitizing range because they simply assume the supplied descriptions are correct (without having to engage in costly exact enumeration of large quantities to verify the description). As such, we view the pairwise comparison paradigm as complementary to elicitation studies.

Nevertheless, the results we report suggest that models of how people describe quantities from images must depend, in part, on the processes by which they perceive those quantities. Models that do not take such perceptual costs into account may be overly permissive, and they may generate descriptions that individuals neither produce nor prefer (Briggs & Harner, 2019). Furthermore, implementing the least effort hypothesis in computational systems for visual scene description will require the ability to model more human-like perception (Kotseruba, Gonzalez, & Tsotsos, 2016). Systems designed to produce natural descriptions of scenes must take into account how people incrementally attend to, perceive, and construct representations of the objects and object clusters depicted in those scenes.

Acknowledgments

This work was supported by a NRC Research Associateship award to HH, a NRL Karles Fellowship awarded to GB. The views expressed in this paper are solely those of the authors and should not be taken to reflect any official policy or position of the United States Government or the Department of Defense. We would also like to thank Danielle Paterno, Kalyan Gupta, and the Knexus Research Corporation for their assistance in supporting these studies. Additionally, we would like to thank Paula Rubio-Fernandez, Andrew Lovett, Greg Trafton, Tony Harrison, Ed Lawson, and the reviewers for helpful feedback in improving the presentation of this work.

References

Anobile, G., Castaldi, E., Maldonado, M. P. A., Burr, D. C., & Arrighi, R. (2020). "groupitizing": a strategy for numerosity estimation. *Scientific Reports (Nature Publisher Group)*, *10*(1).

Barr, D., van Deemter, K., & Fernández, R. (2013). Generation of quantified referring expressions: evidence from experimental data. In *Proceedings of the 14th european workshop on natural language generation* (pp. 157–161).

Briggs, G., & Harner, H. (2019). Generating Quantified Referring Expressions with Perceptual Cost Pruning. In *Proceedings of the 12th International Conference on Natural Language Generation* (pp. 11–18). Tokyo, Japan.

Briggs, G., Harner, H., & Khemlani, S. (2020). Visual grouping and pragmatic constraints in the generation of quantified descriptions. In *Proceedings of the 42nd annual virtual meeting of the cognitive science society* (pp. 1008–1014).

Briggs, G., Wasylyshyn, C., & Bello, P. F. (2019). Elicitation of Quantified Description Under Time Constraints. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (pp. 1436–1442). Montreal, Canada.

Ciccione, L., & Dehaene, S. (2020). Grouping mechanisms in numerosity perception. *Open Mind*, *4*, 102–118.

Cummins, C. (2015). *Constraints on Numerical Expressions* (Vol. 5). Oxford University Press.

Grice, H. P. (1975). Logic and conversation. *1975*, 41–58.

Hesse, C., & Benz, A. (2018). Giving the wrong impression: Strategic use of comparatively modified numerals in a question answering system. In *Proceedings of The Conference on Natural Language Processing (KONVENS)* (pp. 148–157). Vienna, Austria.

Koolen, R., Krahmer, E., & Swerts, M. (2016). How distractor objects trigger referential overspecification: testing the effects of visual clutter and distractor distance. *Cognitive science*, *40*(7), 1617–1647.

Kotseruba, I., Gonzalez, O. J. A., & Tsotsos, J. K. (2016). A review of 40 years of cognitive architecture research: Focus on perception, attention, learning and applications. *arXiv preprint arXiv:1610.08602*, 1–74.

Mandler, G., & Shebo, B. J. (1982). Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General*, *111*, 1–22.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*, 411–419.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*, 89–110.

Rubio-Fernández, P. (2016). How redundant are redundant color adjectives? an efficiency-based analysis of color overspecification. *Frontiers in psychology*, *7*, 153.

Starkey, G. S., & McCandliss, B. D. (2014). The emergence of "groupitizing" in children's numerical cognition. *Journal of Experimental Child Psychology*, *126*, 120–137.

Tarenskeen, S., Broersma, M., & Geurts, B. (2015). Overspecification of color, pattern, and size: salience, absoluteness, and consistency. *Frontiers in Psychology*, *6*, 1703.

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? a limited-capacity preattentive stage in vision. *Psychological Review*, *101*, 80–102.

Viethen, J., van Vessem, T., Goudbeek, M., & Krahmer, E. (2017). Color in reference production: the role of color similarity and color codability. *Cognitive Science*, *41*, 1493–1514.

Wege, T. E., Trezise, K., & Inglis, M. (2021). Finding the subitizing in groupitizing: Evidence for parallel subitizing of dots and groups in grouped arrays. Retrieved from `https://doi.org/10.31234/osf.io/x2ztc`