Some inferences still take time:
Prosody, predictability, and the speed of scalar inferences

Yi Ting Huang[1] & Jesse Snedeker[2]

1. Department of Hearing and Speech Sciences, University of Maryland College Park

2. Department of Psychology, Harvard University

Acknowledgments: This work is supported by grants from NICHD to YH (HD061173) and NSF to JS (BCS-0623845). We thank Noemi Hahn, Kate McCurdy, Elizabeth Casserly, Amanda Worek, and Philip Kim for their help with data collection and ToBI analyses. We also thank Manizeh Khan for comments on an earlier draft and Dan Grodner for sharing the audio files used in his study. This work benefited from the comments of audience members at XPRAG 2009 and the XPRAG master class in 2013. Author address: Department of Hearing and Speech Sciences, 0100 Lefrak Hall, College Park, MD 20742. Email address: ythuang1@umd.edu.

Abstract

Much of the research in experimental pragmatics has focused on how people use the presence of weak scalar terms (like *"some"*) to infer that a stronger alternative (like *"all"*) is false. While earlier work found that comprehenders initially interpret *"some"* without this upper bound, recent work suggests that the inference can be immediate [Grodner, Klein, Carbary, & Tanenhaus (2010). "Some"and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. Cognition, 116, 42-55]. The present paper explores whether rapid inferencing depends on prosody (*"summa"* rather than *"some of"*) or the predictability of referring expressions (e.g., consistently describing subsets with *"some"*). Two eye-tracking experiments examined looks to subsets (2-of-4 socks) and total sets (3-of-3 soccer balls) following *"some"* and found early preferences for subsets in predictable contexts but not in less predictable ones, with no significant effects of quantifier prosody. In our discourse context, predictability had no effect on the perceived naturalness of *"some"* (Experiment 3), but lead to less variability in the verbal encoding of both sets and subsets (Experiment 4). These results suggest that, while scalar inferences are often delayed during comprehension, reference restriction can occur quickly when descriptions of referents can be formulated beforehand.

Keywords: *scalar implicatures, quantifiers, prosody, prediction, semantics, pragmatics*

1. Introduction

Theories of language make a distinction between the linguistically-encoded meaning of an utterance (*semantics*) and how this meaning is enriched (or shifts) depending on context, world knowledge, and the speaker's goals (*pragmatics*). While some such distinction is necessary, the placement of the boundary between semantics and pragmatics can be unclear or counterintuitive. Take for instance the dialogue in (1)

(1) Reporter: Will you answer our questions during the press conference?

Politician: I will answer some of them.

The second statement will typically be interpreted as meaning both that the politician will answer one or more of the questions posed to her and that she will not answer all of them. The later intuition is so strong that it is tempting to assume that the meaning of *"some"* necessarily excludes *all*, but exchanges like (2) demonstrate that this is not the case.

(2) Reporter: Did you address some of the rumors in your book?

Movie star: Yes, I addressed some of them. In fact, I addressed all of them.

In contrast, our knowledge that *"some"* means *one or more*, cannot be overridden or cancelled regardless of the context (3).

(3) Mob Boss: Did you answer any of the officer's questions?

Flunky: *Don't worry! I answered some of them. In fact, I answered none of them!

This pattern has been argued to reflect the distinction between semantically-encoded meaning and pragmatic enrichment (Horn, 1972/1989; Gadzar, 1979). The semantic meaning of *"some"* is lower bounded: it picks out any amount greater than the minimum value on the quantity scale (so any value greater than none, or if the plural is used any value greater than one). Because this content is semantically encoded, it cannot be cancelled, but it can be enriched by

optional pragmatic inferences that depend on conversational goals and our beliefs about speaker's knowledge. For example, in utterances like (1), an upper-bound is added to *"some"* which excludes referents that are compatible with the maximum value on the quantity scale (*all*). This interpretation is motivated by the listeners' expectation that speakers will be *"*as informative as required but no more informative than is required*"* (Grice, 1975). In other words, if the politician had intended to spill every secret, she would have instead uttered something like (4).

(4) Politician: I will answer all of your questions.

Since she did not use this obvious alternative, the listener can infer that there must be some questions that she will not address. Critically, this inference (often called a "scalar implicature") is distinct from the semantics of the expression. Thus, listeners can still access an underlying meaning when the inference is cancelled or never calculated, as in (2).[1]

Psycholinguistic studies of implicature have focused on how these two forms of meaning emerge during language comprehension. The earliest studies measured response times for

[1] The definition of the term "scalar implicature" depends on your theory of the phenomenon. Under some accounts, the implicature in (1) results from the insertion of an operator which negates alternatives (like *all*) and can be embedded in the semantic structure (see e.g., Chierchia, 2004; Chierchia, Fox & Spector, 2013). On these accounts, all the studies that we will be discussing involve the same enrichment process and thus they are all scalar implicatures. However, in other theories, including the classical Gricean account, scalar implicatures are inferences based on entire speech acts and thus truely embedded implicatures are impossible (see e.g., Guerts, 2009; Breheny, Ferguson, and Katsos, 2013a). On these accounts, the utterances in the present study--and in the Huang and Snedeker (2009; 2011) and Grodner studies (2010)--are not scalar implicatures because they are embedded in a definite description. Nevertheless, we will be calling them scalar implicatures because: 1) this framing is consistent with the prior papers under discussion; 2) embedded (local) implicatures do occur (see Chemla & Spector, 2011, Experiment 2) and thus we favor the theories that explain them; and 3) the psycholinguistic studies to date suggest that processing patterns are the same in definite descriptions as they are in standard (upward entailing) contexts (compare Huang & Snedeker, 2009 to Panizza, Chierchia, Huang & Snedeker, 2009 or Grodner et al., 2010 to Breheny et al. 2013a) and thus it would be most parsimonious to treat these as two examples of a single phenomenon. Those who disagree are free to replace this term with their preferred alternative (e.g., "the inference formerly known as SI").

judgments of sentences like *"Some elephants are mammals"* (Rips, 1975; Noveck & Posada, 2003; Bott & Noveck, 2004; Feeney, Scafton, Duckworth, & Handley, 2004; De Neys & Schaeken, 2007). For underinformative statements like these, a false response indicates an interpretation that includes the scalar implicature, while a true response indicates one without it. Bott and Noveck (2004) found that participants who judged the statements to be false took longer than those who judged them to be true. This suggests that scalar implicatures are not calculated immediately during comprehension but instead require time to compute (for related work using speed-accuracy tradeoff method see, Bott, Bailey & Grodner, 2012).

In our prior research, we found further evidence that scalar implicatures are made after some degree of semantic analysis (see Huang & Snedeker, 2009a, henceforth HS, and Huang & Snedeker 2011a). Using the visual-world paradigm, we examined how the interpretation of *"some"* unfolds over the course of an utterance. Participants were presented with instructions like *"Point to the girl that has some of the socks"* while their eye-movements were measured to displays featuring a girl with a subset of one item (2-of-4 socks) and a second girl with a total set of another item (3-of-3 soccer balls). Critically, there was a period of potential ambiguity from the onset of the quantifier to the disambiguation of the noun (*"-ks"*) where the semantics of the quantifier was compatible with both characters. If participants rapidly calculate scalar implicatures, this ambiguity could be resolved immediately since only one of the girls has a proper subset of items. However, after the onset of the quantifier, we found that participants looked equally often at both the subset and total set, leading to slower reference resolution for *"some"* compared to unambiguous terms like *"all"*, *"two"*, and *"three"*. In fact, evidence of a scalar implicature (as indexed by a reliable preference for the subset compared to the total set)

did not emerge until 800ms after quantifier onset. These results strongly suggest that there is a measurable lag between initial semantic processing and the generation of an implicature.

The delay observed in HS is consistent with most of the existing research, including many studies that are cited as evidence for the rapid online calculation of implicatures. For example, Breheny and colleagues (2006) used a reading paradigm where scalar phrases (*"*some of his relatives*"*) were followed by anaphors that referred back to the excluded subset (*"*the rest*"*). They found that reading times at the anaphor were shorter in contexts which encouraged the implicature, suggesting that the upper-bounding implicature had been completed by the time the anaphor was encountered. Critically, in this study, approximately 2000ms passed between the onset of the scalar term and the appearance of the anaphor. Thus these findings are consistent with a model where some semantic analysis of the scalar term occurs prior to the upper-bounding implicature. Similar time lags are present in Bergen and Grodner's (2012) reading study (approximately 1800 to 2400ms) and Nieuwland and colleagues' (2010) ERP experiment (approximately 1300 to 1700ms). In fact, when the time between the onset of the scalar trigger and the anaphor is decreased (from 2400 ms to 900 ms), then the effect at the anaphor disappears (Hartshorne & Snedeker, under review), suggesting that this time is needed for the upper-bounding implicature to be made.[2]

[2] Both Breheny and colleagues (2006) and Bergen and Grodner (2012) also found that the reading time for the scalar trigger (*"some"*) is longer in contexts that support the scalar implicature than in contexts that do not. While this effect is fast, it has not appeared consistenly across studies (Huang & Gordon; Shevaun & Colin, KU folks, Hartshorne & Snedeker, under review; Hartshorne, Snedeker & Kim, in press). Furthermore, its interpretation is ambiguous. The simplest explanation is that the slower reading time at the trigger indicates that the pragmatic processes associated with generating the scalar implicature can begin early but that the resulting upper-bound emerges much later. This would explain why facilitation on the anaphor only emerges when it occurs substantially later than the scalar trigger.

Nevertheless, there is also clear evidence that, under some circumstances, the content of an implicature can become available about as rapidly as the semantically-encoded content of a word. To the best of our knowledge, all of this evidence comes from experiments using the visual world paradigm. The present paper focuses on one of these experiments, Grodner, Klein, Carbary, and Tanenhaus (2010; henceforth GKCT), which uses a paradigm that is very similar to HS. Critically, the discoveries that we make by exploring this particular finding in more depth will also be relevant to understanding the other visual world studies which find evidence for early implicatures (e.g., Breheny, Ferguson & Katsos, 2013ab), as we will explain in the General Discussion.

In the GKCT study, as in the HS studies, sets of objects (e.g., balls and balloons) were divided among characters who differed in gender, and participants were told to select the character who was identified using a quantified noun phrase (e.g., *"Click on the girl who has ___ the balls"*). However, GKCT made two critical changes to the method used in HS. First, the critical instructions were produced with a different prosodic form. Specifically, the quantifier had a reduced vowel rather than an unreduced vowel (*"summa"* instead of *"some-of"*). Second, while HS included trials with number words in addition to scalar quantifiers, GKCT did not. As a result, any given quantity of objects was primarily described with a single term (*"nunna"*, *"summa"*, or *"alla"*). The findings from GKCT contrasted sharply with HS. Within 200ms of the onset of *"summa,"* listeners began abandoning their looks to the total set and shifted their gaze to the subset. In fact, reference restriction for these trials was as rapid as it was for the unambiguous control trials (*"alla"* and *"nunna"*).

The present paper seeks to answer two questions: What accounts for the divergence between these two studies and what can it tell us about the process by which scalar implicatures are made? We consider two broad possibilities.

1. *Prosody*. In GKCT, the motivation for using prosodically-reduced quantifiers was to increase the probability of calculating scalar implicatures by decreasing the ambiguity of *"some."* Specifically, GKCT state that *"some"* can either be a scalar quantifier or a weak determiner (Postal, 1964; Ladusaw, 1994). When it is interpreted as a weak determiner, it may not be construed as being part of the same scale as *"all"* and consequently the implicature may be less frequent. When *"some"* is used in a partitive construction, it must be interpreted as a quantifier and the scalar implicature is more robust. This can be illustrated by minimal pairs like the ones in (4). When the partitive is used (4a), the sentence strongly implies that Ernie ate only a subset of the apples that are present in the context. In contrast, when a bare quantifier is used (4b), there is no explicit domain of quantification and no strong inference that there were any apples left uneaten.

(4a) Ernie: I ate some of the apples.

(4b) Ernie: I ate some apples.

Recent evidence confirms these intuitions. Degen and Tanenhaus (in press) found that participants often accepted the statement *"You got some gumballs"* as an appropriate description of a scene where they got all of the gumballs in the machine. However, given the same context, participants typically rejected the statement *"You got some of the gumballs."*

Thus the use of the phonologically-reduced form (*"summa"*) in GKCT might allow participants to rapidly identify the partitive construction which could facilitate processing relative to HS (where the quantifiers were not reduced). This facilitation could occur in one of

two ways. First, the scalar implicature could be calculated on the fly (from the lower-bounded meaning of *"some"*) but, because the inference is more robust for the partitive, this calculation might occur more rapidly. This view is hard to reconcile with GKCT's findings that the implicature takes no time at all to generate (since *"some"* is given an upper bound as fast as *"all"* is given a lower bound). Nevertheless, it is conceivable these inferences do occur in real time but are so rapid that we cannot detect the lag using current methods. A second possibility (inspired by Levinson, 2000) is that implicatures are retrieved automatically and by default when *"some"* appears in the partitive construction, perhaps because they are stored in the lexicon. The strong version of this hypothesis is clearly false: prior reading time studies have used partitive constructions and found that the implicature is only calculated when the context supports it (Breheny et al., 2006; Bergen & Grodner, 2012). But perhaps the lexical form with the stored upper bound is only accessible in an auditory task where phonological cues are also present.

On either explanation, we would expect slower implicature processing in HS where the prosodic form of the quantifier was consistent with either a partitive construction or a bare quantifier. If the partitive facilitates the rapid online calculation of the implicature, this process might be delayed until the point at which the construction is disambiguated (at *"of"*).[3] Similarly, if phonological reduction allows for the retrieval of a form with a lexicalized upper bound, this route would be inaccessible in the HS study.

[3] Because all of the instructions in the HS study used the partitive construction, participants may have been able to predict that *"some"* was being used as a partitive prior to the preposition. Huang and Snedeker (2011) provide evidence that with minimal experience, participants could predict that the *"some"* was being used as an upper-bounded quantifier prior to the preposition. In an off-line judgment task, participants were given stories and displays, followed by phrase-by-phrase presentation of the critical utterances (*"Point-to-the-girl-that-has-some"* | *"of-the"* | *"ice-cream"* | *"sandwiches"*). After each phrase, participants were asked to choose whether the likely referent was a girl with a subset (2-of-4 ice cream sandwiches) or total set (3-of-3 ice cream cones). After just the first segment, they showed a reliable preference for the subset.

2. *Number words*. A second critical difference between the two studies was the range of descriptions that were used across trials. In HS, half of the instructions used number words (*"Point to the girl that has two/three of the socks"*) and half used scalar quantifiers (*"Point to the girl that has some/all of the socks"*). As a result, participants in the HS study could not reliably predict what lexical label would be used for a particular set (e.g., *"some"* or *"two"*) or what conceptual encoding would be most relevant (subset or exact numerosity). In contrast, in GKCT, the target was identified using a scalar quantifier on 85% of the trials *("Click on the girl who has nunna/summa/alla the balls"*). The remaining trials used simple definites *("Click on the girl who has the balloons"*). As a result, the lexical label that would be used for any given set was partially predictable before the utterance began and only one conceptualization of quantity was ever relevant. To be clear, it was not the case that listeners could predict *which* character would be the target before the sentence began (since the probabilities of referring to the characters with the subsets, total sets, and empty sets were equated). However, listeners could potentially predict how any given set in the scene would be described in the instructions if that character was the target: total sets were always be labeled with *"alla"*, empty sets were always labeled with *"nunna"*, and subsets were labeled with *"summa"* on the majority of the trials (70.6% *"summa the X's"*, 29.4% *"the X's"*).

The presence of number words could affect the online interpretation of scalar quantifiers in one of two ways.

One possibility is that predictable stimuli allowed participants to encode the display in a manner that allows them to bypass the scalar implicature during comprehension and directly map the quantifier *"some"* to the subset display (Huang & Snedeker, 2009b; Huang et al., 2010). For example, prior to the instruction, listeners who view the display might conceive of the girl with a

subset of the balls as *"the girl who has some of the balls"* and the girl with the total set of balloons as *"the girl who has all of the balloons."* This kind of spontaneous prediction of verbal descriptions is broadly consistent with theories that posit a tight coupling between language comprehension and language production (Pickering & Garrod, 2013; Dell & Chang, 2014). Critically, predictions of this kind could result in what look like instantaneous implicatures. Specifically, as the instructions unfold, listeners could compare the external input to their own internal description of a referent, sticking with a referent if the descriptions match, or ruling it out if the descriptions diverge. This mechanism could allow listeners to immediately link *"some"* to the subset and rule out the total set, without having to first retrieve its meaning and then generate an upper bound through an inference. On this account, the scalar implicature is instantaneous because listeners (like speakers) tend to encode visual contexts in linguistically-relevant ways. In contrast to the GKCT, the design of HS presented participants with two equally likely conceptualizations of each set. This may have discouraged the use of predictive encoding during comprehension (since it is less useful in these circumstances) or blocked the use of previously-generated descriptions (since a divergence between the internal description and the external input would no longer rule out the possibility that they describe the same referent).

GKCT and their colleagues have proposed an alternative explanation for how the use of number words could affect the interpretation of *"some"* (Grodner et al., 2010; Degen & Tanenhaus, in press). Specifically, they suggest that the presence of numbers in the HS study focuses attention to the exact numerosity of the set, making it unnatural to describe the subset displays with *"some"*. On this account, the delay in interpreting the *"some"* utterances in the HS study reflects the fact that participants do not consider them to be good descriptions of either the subset or total set. We will discuss the details of this proposal in the General Discussion.

Critically, for our present purposes, both accounts predict that delays in the interpretation of *"some"* should occur whenever number words are used in the instructions.

The present study seeks to determine whether the differences between HS and GCKT are due to prosodic form of the quantifier or the presence of number trials. Experiment 1 tests the prosodic hypothesis by exploring the interpretation of the phonologically-reduced forms (*"summa"*) in a context where the subset (girl with 2-of-4 socks) was labeled with *both* scalar quantifiers and number words. Experiment 2 pits the prosodic hypothesis against the number hypothesis by factorially manipulating both variables in a between participants design. The results of these experiments demonstrate that prosody has no effect on the timing of the upper-bounding implicature. The presence of number trials, however, has a strong effect: when number words are absent, the upper bound of *"some"* is available as quickly as the lower bound of *"all"*. The last two experiments use a ratings task (Experiment 3) and a production task (Experiment 4) to explore why the presence of number words affects the interpretation of *"some."*

2. Experiment 1

This experiment used the paradigm developed in the HS study, but with phonologically-reduced quantifiers similar to those used in GKCT. On the critical trials, participants were asked to *"Click on the girl that has summa the socks"* while their eye-movements were measured in two different contexts. On the 2-referent trials, participants saw the subset paired with a total set (girl with 3-of-3 soccer balls). In this context, the referent of the scalar quantifier is semantically ambiguous but pragmatically unambiguous. On the 1-referent trials, participants saw the subset paired with an empty set (girl with 0-of-3 soccer balls). In this context, the referent of the scalar quantifier can be distinguished by semantics alone. Finally, we included control trials asking for

the competing set (the total or empty set) using unambiguous quantifiers (*"alla" or "nunna"*) and filler trials asking for the total set, subset, or empty set using alternate descriptions (*"two of"* for the subset, *"three of"* for the total set, and *"didn't get"* for the empty set). As in the HS study, these trials decreased the predictability of the quantity-to-quantifier mapping since each type of set was labeled in two ways each with equal frequency.

These materials allow us to explore two explanations for the rapid preference for the subset found by GKCT. If scalar implicatures are immediate for phonologically-reduced expressions, then participants should rapidly make the upper-bounding implicature in *"summa"* trials. As a result, reference resolution should be as fast for the 2-referent trials as it is for the 1-referent trials. However, if the rapid implicature in GCKT reflects other differences between the studies, such as the predictability of the verbal label, then participants should continue to show a delay in calculating the upper bound of *"some,"* as they did in HS. This would result in slower reference resolution in the 2-referent *"summa"* trials than in the 1-referent *"summa"* trials.

2.1. Methods

2.1.1. Participants

Forty English-speaking undergraduate students at Harvard University participated in this study. They received either course credit or $5 for their participation.

2.1.2. Procedure

Participants sat in front of a computer display and their eye movements to the screen were measured using a Tobii T60 eye-tracker. At the beginning of the study, participants were told that they would hear and see a series of stories about two boys and two girls (Craig, Pat, Judy, and Cheryl). Each character had a fixed position on the screen, which was divided into quadrants. Every trial consisted of a story followed by a critical utterance. In each story, two

sets of objects were distributed and pictures of these objects appeared next the character who received them. The critical utterances instructed participants to select a one of the characters by clicking on it with the mouse. Once the participant did this, the trial ended. Participants were then were instructed to click the mouse again to proceed to the next trial. To ensure that participants understood the task, two practice trials without quantifiers were presented prior to the test trials (e.g., *"Click on the girl that got the apple"*).

2.1.3. Materials

Each participant was assigned to one of two conditions which varied in terms of the Display Type that was used. Participants in the 1-referent condition saw critical trials in which a subset (2-of-4 socks) was contrasted with an empty set (0-of-3 soccer balls). For participants in the 2-referent condition, a subset was contrasted with a total set (3-of-3 soccer balls). These displays allowed us to compare interpretations of *"summa"* that required an implicature and those that did not. Participants in both conditions saw sentences with two different Quantifier Types: the critical *"summa"* trials (which always picked out the subset) and the control trials which picked out the other character using a quantifier. In the 1-referent display, this alternative was labeled with *"nunna."* In the 2-referent display, it was labeled with *"alla"*.

Figure 1 provides examples of the visual displays that were used. Each display included the four characters who appeared in the same location throughout (clockwise from the upper-left quadrant: Craig, Judy, Cheryl, and Pat). This arrangement ensured that the vertically-adjacent characters matched in gender while the horizontally-adjacent characters did not. On each trial, participants heard a story like (5), in which two types of objects were introduced and distributed among the boy-girl pairs.

(5) The boys and girls on the soccer team were getting socks and soccer balls from the coach. The coach gave socks to Judy and socks to Craig (*two socks appear next to the girl on the upper-right and two socks next to the boy on the upper-left*). The coach knew that Pat was already a very good soccer player but he thought that Cheryl needed a lot of practice (*nothing appears next to the boy on the lower-left and three soccer balls next to the girl on the lower-right*).

These stories always involved one set of four items that was split evenly between a horizontally-adjacent boy-girl pair and another set of three items which was given to one member of the other pair. By introducing the objects as part of a single large set and then dividing that set among the characters, these stories established a clear domain of quantification for the critical utterances. For example, after a story like (5), *"alla the soccer balls"* most naturally refers to all the soccer balls that the coach had, rather than all of the soccer balls in the known universe or all of the soccer balls that Cheryl has. These stories also ensured that participants knew the nouns that we would use for the objects, which were always referred to with definite noun phrases (*"the socks"*) or bare plurals (*"socks"*) to avoid linking the sets to the numbers or quantifiers used in the critical utterances. In separate off-line judgment tasks, these stories and displays were found to successfully establish the expectations that: (a) quantifiers would refer specifically to the sets in the display, (b) objects would be identified by basic-level labels, and (c) *"summa"* would be interpreted with a scalar implicature (see Huang and Snedeker 2009a; 2011a for more details).

INSERT FIGURE 1 ABOUT HERE

For each story, critical utterances were created like those in (6).

(6) Click on the girl that got <u>summa/nunna/alla</u> the socks.

The gender of the character varied across trials and was linked to the content of the story. In the 2-referent trials, if the set of three objects had been given to a girl, then a girl was requested. In contrast, in the 1-referent trials, if the set of three objects had been given to a girl, then a boy was requested. The names of the two kinds of objects always shared a phonological onset (socks and soccer balls), creating a brief period of ambiguity during which the identity of this noun was uncertain. In our analyses, the character who was requested is the Target (the girl with socks) while the one who matched in gender is the Distractor (the girl with soccer balls or nothing). Eight critical items were generated by creating four versions of each base item which were then distributed across four presentation lists such that each list contained four items in each condition and each base item appeared just once in every list.

Finally, eight additional filler trials using number words and descriptions were included to ensure that the verbal label for each type sets could not be reliably predicted. In 1-referent condition, the Target with the empty set was described in the filler trials as *"the girl that didn't get the socks"*. In the 2-referent condition, the total set was described in the filler trials as *"three of the socks."* In both conditions, the subset was described in the filler trials as *"two of the socks."* Because the stories and displays preceding the instructions were of the same format for the critical and filler trials, participants could not predict the quantifier that would be used prior to hearing it. The stimuli for all the experiments are provided in Appendices A and B.

Target utterances were recorded by a female actor and the sound files were edited to ensure that the lengths of two regions were equated across the four conditions: 1) the region from sentence onset to the gender cue (*"Click on the"*) and 2) the region from the onset of the gender cue to the onset of the quantifier (*"girl that got"*). To ensure that the sentences were produced in the desired fashion, a trained research assistant coded the utterances for the *"summa"* trials using

the ToBI annotation system (Beckman & Hirschberg, 1994) and compared them with trials from GKCT and HS. These analyses revealed that utterances like *"some of the socks"* versus *"summa the socks"* are distinguished based on whether the consonant at the end of *"of"* is articulated or not. This consonant was not articulated in any of the instructions in the current experiment or in GKCT. As a result, there was no word level break between *"summa"* and *"the."* In contrast, the instructions used in HS consistently had a word-level break (a 1) at this juncture.[4]

2.2. Results

We examined the proportion of participants' fixations to the Target character on two different time scales. Our first analysis examined a coarse-grained measure of participants' fixations as the target utterance unfolded and examined five broad time windows:

1. Baseline phase: This 500ms period begins at the onset of the instruction and ends just before the onset of the gender cue (*"Click on the"*). This region provides a baseline measure of looks to the display before any gender or quantifier information.
2. Gender phase: This 650ms period begins at the onset of the gender cue and ends just before the onset of the quantifier (*"girl that got"*). This region provides a direct comparison of looks to the Target and Distractor before any quantifier information. Here we predict that fixations will shift towards the side of the display with characters that match the specified gender.
3. Quantifier phase: This 700ms period begins at the onset of the quantifier and ends just before the onset of the disambiguating phoneme (*"summa/nunna/alla the soc-"*). In this region, we expect that participants will begin using information about the quantifier to

[4] Across the three stimulus sets (HS, GKCT, and the present study), there were no differences at the juncture between *"some"* and *"of."* Utterances were always produced with no break (a 0) in this location. This is because *"of"* is a clitic that begins in a vowel and since English is a language that prefers to build right-headed feet, *"some-of"* makes a perfect metrical foot.

look to the Target. By comparing the four trial types, we can determine whether there are delays in reference restriction for contrasts that require a scalar implicature versus those that can be resolved based on the semantics of the quantifier. We expect Target looks to rapidly increase when participants hear *"nunna"* and *"alla"*, or *"summa"* in the 1-referent trials. If scalar implicatures are calculated immediately in the presence of an early phonological cue to the partitive, we should also see a preference for the Target following *"summa"* in the 2-referent trials. However, if scalar implicatures are preceded by a period of initial semantic analysis, then participants should continue looking equally to both the Target and Distractor for the *"summa"* 2-referent trials.

4. Disambiguation phase: This 400ms period begins at the onset of the disambiguating phoneme and ends at the offset of the command (*"-ks"*). This region unambiguously resolves the correct referent by picking out the item that he or she possesses. Here fixations should be primarily on the Target with relatively few looks the Distractor.
5. End phase: This period begins at the end of the sentence and continues through a 500ms window. Again we predict that across all trials, participants would be looking at the Target prior to initiating their selection.

In the analyses of the large time windows, the onset of each region is shifted 200ms after the relevant marker in the speech stream to account for the time it would take to program a saccadic eye movement (Allopenna, Magnuson, & Tanenhaus, 1998; Matin, Shao, & Boff, 1993).

Our primary dependent measure was preference for the Target over the Distractor. This was calculated as the number of samples (for a given trial and a given window) in which the participant looked at the Target minus the number of samples in which they looked at the Distractor. If this number was positive, Target preference was 1. If it was negative, then Target

preference was 0. If participants looked at neither object or at both objects equally during a time window, this sample was excluded from the analysis. This binary variable best captures the underlying distribution of our data. While our eye-tracker samples eye gaze 60 times a second, people typically make only a couple of saccades in a second. Consequently, in a short time window, most participants will only fixate one of the possible objects, and thus any measure of fixation proportion within that window is essentially binary. We analyzed these data in a logistic mixed-effects model using the lme4 software package in R (Bates, 2007). Subjects and items were modeled as simultaneous random effects on the intercept only.[5]

Figure 2 illustrates that prior to the onset of the quantifier, looks to the Target were around chance across all conditions. There were no reliable effects during the Baseline and Gender phases (all *p's* > .30). Critically, during the Quantifier phase, a preference for the Target emerges in three of the four trial types. In the 1-referent display condition, Target looks increased following both *"nunna"* (82%) and *"summa"* (82%). However, in the 2-referent display condition, Target looks increased following *"alla"* (66%) but not following *"summa"* (43%). This led to significant main effect of Display Type ($z = 4.36$, $p < .001$), a marginal effect of Quantifier Type ($z = 1.66$, $p < .10$), as well as a significant interaction between the two ($z = 2.10$, $p < .05$). Planned comparisons within each Display condition revealed no difference in Target looking between *"summa"* and *"nunna"* in the 1-referent condition ($z = 0.26$, $p > .70$) but a robust difference between *"summa"* and *"alla"* in the 2-referent condition ($z = 3.18$, $p < .01$). This demonstrates that listeners were quick to disambiguate referents that could be identified solely on the basis of the semantics of the quantifier. However, when a pragmatic implicature

[5] Final models were selected by first including all main effects and interactions and then removing predictors until the fit of the smaller model was not significantly worse than the fit of the full model ($p > .05$). Initial models also included random slopes, but in no case did this result in a significant improvement in model fit. Thus, they were excluded from further analyses.

was required (as in the case of 2-referent *"summa"*), reference resolution was delayed. This strongly suggests that the scalar implicature is not immediately available, even when the phonological form of the quantifier signals that the partitive construction is being used.

INSERT FIGURE 2 ABOUT HERE

During the Disambiguation phase, looks to the Target increased rapidly for the 2-referent *"summa"* trials (72%) suggesting that the phonological information allowed participants to close in on the correct referent. However, Target looks in these trials continued to lag behind those in the *"nunna"* (92%), *"alla"* (89%), and 1-referent *"summa"* (91%) trials. This led to significant main effects of Display Type ($z = 2.63$, $p < .01$), no effect of Quantifier Type ($z = 1.04$, $p > .20$) and a marginal interaction between the two ($z = 1.69$, $p < .10$). Planned comparisons within the 1-referent condition again revealed no difference in Target looks between *"summa"* and *"nunna"* ($z = 0.22$, $p > .80$). In contrast in the 2-referent condition, there were significantly fewer Target looks for *"summa"* than *"alla"* ($z = 2.51$, $p < .05$). Finally, during the End phase, after the offset of the instructions, participants across all conditions closed in on the Target, resulting in no reliable main effects or interactions (all p's $> .90$).

Our second analysis focused on the period immediately after the quantifier onset in order to explore the critical interaction in greater detail. Target preference was calculated for each 100ms interval following the onset of the quantifier. These time windows were not offset by 200ms and each window included all samples from the labeled time point to the moment prior to the onset of the next interval (e.g., the 100ms window included eye data collected between 100ms after quantifier onset to 199ms after quantifier onset). These analyses revealed a significant Display by Quantifier Type interaction that emerged in the 200ms time window ($z = 3.08$, $p < .01$) and continued through the 800ms window ($z = 1.95$, $p < .05$). These results indicate that despite the

presence of phonologically-reduced quantifiers (*"summa"* rather than *"some-of"*), the generation of the scalar implicature was substantially delayed.

2.3. Discussion

In Experiment 1, we found increases in looks to the Target shortly after the onset of *"nunna"*, *"alla"*, and *"summa"* in a 1-referent display context. In contrast, during the 2-referent *"summa"* trials, participants continued to look equally at the Distractor (total set) and the Target (subset). These results suggest that, when lexical semantics is sufficient to identity the correct referent, disambiguation is quite rapid. However, when semantic analysis is not sufficient and identification requires an additional pragmatic inference, reference resolution is substantially delayed. This pattern replicates prior findings by Huang and Snedeker (2009a) and suggests that scalar implicatures are not immediately available during real-time comprehension. Critically, these findings suggest that prosody alone cannot account for the different data patterns observed in the HS and GKCT studies. If the phonologically-reduced form was directly linked to a meaning that licensed the scalar implicature, then reference resolution in the 2-referent *"summa"* trials should have been as rapid as it was in the 1-referent *"summa"* trials. Instead these results indicate that providing prosodic cues which signal the partitive construction is not enough to create rapid, let alone instantaneous, scalar implicatures.

These results are consistent with the hypothesis that the presence of number trials may affect whether the scalar implicature in delayed. Experiment 1, like HS but unlike GKCT, included number trials (in the filler item) and found a delay in calculating the upper-bound of *"some."* As we noted in the Introduction, two possibilities have been discussed for why the inclusion of numbers has the effect that it does. First, we have argued that having predictable verbal labels for each quantity allows participants to encode the visual context in linguistically

relevant ways (Huang & Snedeker, 2009b; Huang et al., 2010). By matching the verbal input to these descriptions, listeners could map the scalar quantifier *"some"* directly to the character with a subset of items. Second, Grodner and colleagues have argued that including numbers in the context makes the *"some"* sentences unnatural, resulting in a delay in processing (Grodner et al., 2010; Degen & Tanenhaus, in press).

Experiment 2 directly contrasts the role of prosody and predictability by factorially manipulating these variables in a between-subjects design. Two levels of predictability were contrasted by varying the types of labels for the sets in the filler trials (number words vs. scalar quantifiers) and two levels of prosody were contrasted by varying the articulation of the critical expression (*"some-of"* vs. *"summa"*). By having the same auditory instructions recorded by the same speaker, we can directly control for irrelevant differences arising across studies (e.g., the exact words in the utterance and speaker variation) and more accurately compare the relevant prosodic differences between *"some-of"* and *"summa."*[6] We will evaluate the relative facilitation of these cues by comparing its effects on the reference restriction of the semantically-ambiguous *"some"* to the semantically-unambiguous *"all."* Since interpretation of *"all"* is rapid by all accounts, this measure will provide an indication of the relative delay of *"some"* across levels of prosody and predictability.

3. Experiment 2

3.1. Methods

3.1.1. Participants

[6] Experiment 2 also explored two additional factors that might have contributed to differences from GKCT. First, we increased the overall number of trials, from four critical *"some"* trials to eight. Second, we included filler trials that referred to characters on both sides of the display, sometimes using scalar quantifiers. Both these features may have increased the likelihood that participants encoded the subset as *"some"* over the course of the study.

Eighty English-speaking undergraduate students at Harvard University participated in this study. They received course credit or $5 for their participation.

3.1.2. Procedure

The procedure was identical to Experiment 1.

3.1.3. Materials

The eight critical conditions were derived from cells of a 2 x 2 x 2 design. The first factor, Quantifier Strength, was manipulated within-subjects and distinguished the weaker term (*"some"*) from the stronger one (*"all"*).[7] The second factor, Prosody, varied whether articulation of the quantifier was phonologically-reduced (*"summa"*) or unreduced (*"some-of"*). This was manipulated between subjects to avoid any confusion resulting from comparisons of the prosodic variation (e.g., the possibility that the two forms would be interpreted contrastively). The third factor, Predictability, varied whether the sets were labeled in the critical instructions in a uniform fashion or variable fashion. This was manipulated by including fillers that used either scalar terms (*"some"/"all"*) to create uniformity, or numbers (*"two"/ "three"*) to create variability. This variable was manipulated between subjects since the type of filler trials was intended to influence participants' perceptions of the experimental interaction as a whole.

As in Experiment 1, every trial began with a story about two sets of objects distributed among the two pairs of characters, which was followed by an instruction asking for one of the characters. Sixteen critical trials asked for subsets and total sets on the 2-referent side of the display using *"some"* and *"all."* These items were generated by creating eight versions of each base item which were then distributed across eight presentation lists such that each list contained

[7] For the sake of clarity, quantifier names (i.e., *"some"*, *"all"*) in this experiment will refer to variations across the strength of these terms. In contrast, variations in prosody will be referred to by their phonological differences (i.e., reduced, unreduced).

eight items in each condition and that each base item appeared just once in every list. Two types of filler trials were also randomly interspersed. First, 16 filler trials asked for subsets and total sets on the 2-referent side. Depending on the condition, these trials used terms from either the quantifier scale (scalar filler: *"some"/"all"*) or number scale (number filler: *"two"/"three"*). Second, to ensure that all four characters in the display were equally likely to be potential referents in this task, 32 filler trials asked for subsets and empty sets on the 1-referent side. The terms used in these trials also varied depending on the condition. In the scalar filler trials, referents were requested using quantifiers (*"some"/"none"*). In the number filler trials, referents were requested using quantifiers, number words, and descriptions (*"some"/"two"/ "none"/ "didn't get"*). The prosody of these fillers varied depending on the condition and matched those of the critical trials. Critically, the combination of these trials ensured that a subset would be referred to with *"some"* 100% of the time in the scalar filler condition but only 50% of the time in the number filler condition.

A trained research assistant ToBI coded the critical instructions for the *"some"* trials across the two levels of Prosody. These analyses confirmed there was never any break between *"summa"* and *"the"* in the reduced-phrasing condition (0 in ToBI) but there was a consistent word-level break (1 break) in the full-phrasing condition.

3.2. Results

Eye-movements were analyzed using the same dependent measures and analytic strategy as in Experiment 1. To compare the effects of Predictability, we separately analyzed the number filler and scalar filler conditions and examined how the time-course of Target preference in each varied across levels of Quantifier Strength and Prosody. First, we examined these changes across the five coarse-grained time windows. The lengths of these regions were as follows:

Baseline (700ms), Gender (600ms), Quantifier (600ms), Disambiguation (500ms), and End phase (600ms). During the Baseline and Gender phases, we found no effects of Quantifier Strength, Prosody, or interaction between the two (all *p's* > .20).

However, differences across conditions emerged at the Quantifier phase (Figure 3). In the number filler condition, Target preference increased following *"all"* (61%) but not *"some"* (51%). This led to a significant main effect of Quantifier Strength (z = 2.82, $p < .01$), with no additional effect of or interaction with Prosody (all *p's* > .80). These results are consistent with patterns found in Experiment 1 and suggest that reference resolution for *"some"* is delayed when implicatures are required for interpretation. In contrast, in the scalar filler condition, a different pattern emerged. Target preference was slightly greater when quantifiers were phonologically-reduced (60%) compared to phonologically-unreduced (51%). This led to a marginal effect of Prosody (z = 1.68, $p < .10$), suggesting that participants were sensitive to the prosodic differences across the instructions. Critically, there was no additional effect of or interaction with Quantifier Strength (all *p's* > .30). This absence of a difference between *"some"* and *"all"* in the scalar filler condition demonstrates that when the subset is consistently labeled with scalar terms, the interpretive delays associated with generating a pragmatic implicature vanish.

INSERT FIGURE 3 ABOUT HERE

Curiously, during the Disambiguation phase, Target preference in the *"some"* trials lagged behind those of the *"all"* trials in the number (74% vs. 88%; z = 4.08, $p < .001$) *and* scalar filler conditions (80% vs. 87%; z = 2.48, $p < .05$), leading to a main effect of Quantifier Strength in both cases. These results suggest that even in a context where verbal encoding can facilitate rapid reference restriction for "*some*," evidence of bottom-up interpretation still persists. No additional effect of or interaction with Prosody was found (all *p's* > .15). Finally, during the End

phase, participants across all conditions closed in on the Target, resulting in no reliable main effects or interactions in both number and scalar filler conditions (all p's > .20).

To examine the time-course of Target preferences in greater detail, we again focused on 100ms intervals following the onset of the quantifier. In the number filler condition, a significant main effect of Quantifier Strength emerged in 200ms time window ($z = 2.67$, $p < .01$) and continued through the 1300ms window ($z = 3.05$, $p < .01$). In contrast, in the scalar filler condition, no reliable effects of Quantifier Strength were found in any time window (all p's > .15). These results suggest that when labels cannot be anticipated prior to the instructions, interpretation of *"some"* is delayed compared to *"all."* However, when these labels are highly predictable, reference restriction for *"some"* is as quick as *"all."* No additional effects of or interactions with Prosody were found in either condition (all p's > .20).

Finally, given the extended length of Experiment 2, we were interested in how quickly predictability effects emerge over the course of the study. To explore this, we separately analyzed first- and second-half trials during the Quantifier and Disambiguation phases. In the number filler condition, for the first-half trials, Target preference following *"some"* lagged behind *"all"* during the Quantifier phase ($z = 2.06$, $p < .05$). For the second half-trials, this effect was marginal ($z = 1.86$, $p < .10$). In the scalar filler condition, there is no effect of Quantifier Strength in either the first- or second-half trials (all p's > .80). The fact that both of these patterns are in place in the first half of the study indicates that listeners can quickly learn to construe scenes in the way that is most relevant given the communicative contexts. In other words, these predictability effects do not appear to be the product of a hard-learned experimentally-induced strategy but instead seem to reflect the intrinsic flexibility of language processing.

Nevertheless, analysis of the Disambiguation phase suggests that even when verbal encoding promotes rapid disambiguation, bottom-up interpretation via lexical meaning and scalar implicature still persist. During the first-half trials, *"some"* was delayed compared to *"all"* in not only the number filler condition ($z = 2.25$, $p < .05$) but in the scalar filler condition as well ($z = 2.83$, $p < .01$). However, while effects of Quantifier Strength continued into the second-half trials of the number filler condition ($z = 3.49$, $p < .01$), they disappeared in the scalar filler condition ($p > .80$). This suggests that additional experience in predictable contexts can dampen even late-emerging effects of bottom-up processing. No effects of or interactions with Prosody were found in either condition or time window (all p's > .20).

3.3. Discussion

In Experiment 2, we found that participants' ability to restrict reference for *"some"* was influenced by the predictability of how the subset was encoded. When different labels were used over the course of the experiment, sentences with *"all"* were interpreted more rapidly than those with *"some."* However, when the sets were consistently labeled with scalar quantifiers, participants initially disambiguated *"some"* as robustly as they disambiguated *"all."* Phonological reduction appeared to have no effect on the scalar implicature: *"some-of"* was processed as rapidly, or as slowly, as *"summa."* Thus we have a clear answer to our initial question: the difference in findings between HS and GKCT is attributable to differences in the range of commands that were used in the two studies. By using only scalar quantifiers and no numbers, GKCT created conditions under which the upper bound of *"some"* restricted reference as rapidly as the lower bound of *"all."* These predictability effects, however, could be attributable to either verbal encoding or the evaluation of naturalness.

One feature of this data seems to uniquely support the verbal-encoding hypothesis. In the scalar filler condition, we found that no reliable difference in Target looks immediately after the onset of *"some"* and *"all"* (Quantifier phase), but there was a reliable difference in a later time region (Disambiguation phase). The verbal-encoding hypothesis predicts that disambiguation for *"some"* can happen via two mechanisms: 1) a rapid mapping from the form directly to the referent facilitated by verbal encoding; 2) a slower process of calculating the upper-bound from the lexical meaning via a contrastive scalar implicature. On this hypothesis, we might expect that that the earliest shifts to the Target will typically reflect verbal encoding (resulting in no difference between *"some"* and *"all"*). In contrast, later shifts will typically occur when verbal encoding was not employed, resulting in earlier disambiguation for *"all"* than *"some"* in later time windows.

To the best of our knowledge there is just one concrete piece of evidence the supports the naturalness hypothesis over the verbal-encoding hypothesis. Appendix B of GKCT reports an offline ratings study where participants were presented with the displays from the eye-tracking study, and were asked to judge the naturalness of different descriptions (see Degen and Tanenhaus (in press) for similar ratings). Half of the participants heard only *"some," "all,"* and *"none"* sentences, while the others heard number sentences as well. Those in the number condition found *"two"* to be a more natural description of the subset than *"some."* Furthermore, while *"none"* and *"all"* were marginally more natural in the presence of numbers, judgments of *"some"* were marginally higher when number trials were absent ($p = .09$). According to GKCT, this demonstrates that participants' preference for the numerical descriptions of subsets interferes with their ability to construe these as instances of *"some,"* leading to the observed delays.

We see two limitations of this finding. First, it fails to explain why the presence of number trials does not result in a slow-down in the interpretation of *"some"* when the Distractor is an empty set (the 1-referent contexts of Experiment 1). *"Some"* should have been an equally unnatural description of the subset in this context, but we found that reference resolution was rapid. Second, it is not clear that the ratings obtained by GKCT reflect the naturalness of these descriptions in *our* task. In the GKCT procedure, each trial involved a display paired with a numerical description (*"There are four balls, four planets, and four balloons"*) followed by the critical instruction (*"Click on the girl who has summa the balls"*). In contrast, in our procedure, displays were always introduced with a story describing sets of objects distributed among characters. Since these richer discourse contexts highlight the division of the sets, they may increase the naturalness of the scalar descriptions.

To explore this possibility, we collected naturalness ratings on our materials and explored the effect of our stories on these ratings. In Experiment 3, we manipulated predictability (scalar trials only vs. scalar/number trials mixed) and the discourse context (with stories vs. without stories) in a between-subjects design. Critically, the naturalness hypothesis predicts the following patterns in participants' naturalness ratings of *"some"* and *"all."* First, the naturalness of *"some"* as a description of the subset will decrease when numbers are used, but the naturalness of *"all"* as a description of the total set should not be affected. Second, numbers will be rated as a more natural description of the subset than *"some,"* but they will not be rated as a more natural description of the total set than *"all."*

4. Experiment 3

4.1. Methods

4.1.1. Participants

Sixty-four English-speaking undergraduate students at the University of North Carolina at Chapel Hill participated in this study. They received course credit or $10 for their participation.

4.1.2. Procedure

The procedure was adopted from the naturalness ratings task described in Appendix B of GKCT, with minor modifications. Participants sat in front of a computer display. At the beginning of the study, they were told that they would see a display depicting two pairs of characters. For half the participants, these displays would be accompanied with stories describing the scene and for the other half, no stories would be provided. Afterwards one of the characters in the display was highlighted with a red box and referred to with an instruction. GKCT presented these commands in written form. However, to make this task more similar to the experience of our eye-tracking participants, we presented our commands aurally. Participants were told to rate how naturally the utterance identified the character, using a scale that ranged from *"1"* for very unnatural to *"7"* for very natural.

4.1.3. Materials

The stories, displays, and instructions came from Experiment 1 of Huang and Snedeker (2009a). There were four conditions as representing the four cells of 2 x 2 between participants design. As in Experiment 2, the first factor, Predictability, reflected whether the critical instructions labeled the sets in a uniform fashion using scalar fillers (*"some"/"all"*) or in a variable fashion using number fillers (*"two"/"three"*). The second factor, Discourse Context, indicates whether a story was presented for each display or not. In the story condition, the stories and displays unfolded in the same way as in the eye-tracking studies while in the no story condition, the characters and sets of objects were presented in their final form. Both Predictability and Discourse Context were manipulated between subjects since both factors were

intended to influence participants' perceptions of the experimental interaction as a whole. Within each of these four conditions, two additional variables were manipulated. The first variable, Trial Type, distinguished the critical *"some"* and *"all"* trials from the filler trials used in the Predictability manipulation (*"some"/"all"* or *"two"/"three"*). The second variable, Quantifier Strength, distinguished the weaker term (*"some"*) from the stronger one (*"all"*). Both were manipulated within subjects and counterbalanced across four presentation lists.

4.2. Results and Discussion

Our dependent measure was the mean naturalness rating of the utterances. Figure 4 illustrates that instructions were generally considered felicitous, with average ratings across all conditions exceeding the midpoint. To investigate patterns in greater detail, we analyzed the effects of Predictability, Discourse Context, and Quantifier Strength on the natural log of the ratings using a series of linear mixed-effects models. Subjects and items were included as simultaneous random effects on the intercept only. Significance tests for fixed effects were estimated using a Monte Carlo Markov Chain procedure (MCMC).

INSERT FIGURE 4 ABOUT HERE

We conducted our analyses in three parts. First, we focused on judgments in the no story condition and tested the key predictions of the naturalness hypothesis. Recall that in the absence of a richer discourse, GKCT found that number fillers increased the naturalness of *"all"* but decreased the naturalness of *"some."* In contrast, our analyses revealed no significant main effects Quantifier Strength and Predictability and no interaction between the two (all p's > .15). Next, in the number condition we compared the ratings of *"two"* and *"some."* Here we replicated GKCT's findings that *"two"* was a more natural description of the subset compared to *"some"* (6.3 vs. 5.3; t = 5.00, $p < .001$). Finally, we focused on changes in the naturalness of

"some" in the presence and absence of number fillers. We replicated GKCT's findings that *"some"* was marginally less natural when labels are variable compared to uniform (5.3 vs. 5.9; t = 1.54, $p < .10$). These latter two findings are consistent with predictions made by the naturalness hypothesis.

Second, we examined whether the same patterns emerged in the story condition. Since these discourse contexts closely approximate those used in Experiment 2, analysis of these ratings allows us to assess whether naturalness alone can explain delays in interpreting *"some."* The analysis of Quantifier Strength and Predictability again revealed no significant main effects or interactions (all $p's > .30$). However, in contrast with the no story condition, *"some"* was now judged to be as natural as *"two"* (5.6 vs. 5.8; $p > .50$). Similarly, ratings of *"some"* were no different in the presence or absence of number fillers (5.6 vs. 5.6; $p > .70$). Thus none of the predictions of the naturalness hypothesis held up in the story condition.

Finally, we directly compared how the naturalness of total set and subset descriptions changed depending on the discourse context. This analysis directly compared numbers and scalars and thus included only the low predictability conditions. Analysis of total sets revealed that regardless of whether stories were used, *"three"* was a more natural description than *"all"* (6.1 vs 5.6). This led to a main effect of Scale Type (t = 3.14, $p < .01$), with no additional effect of or interaction with Discourse Context (all $p's > .50$). Analysis of subsets also revealed that *"two"* was a more natural description than *"some"* (6.1 vs 5.4; t = 3.15, $p < .01$). Critically, this main effect of Scale Type was qualified by an additional interaction with Discourse Context (t = 2.26, $p < .05$). While ratings for *"two"* were higher in the no story condition compared to the story condition (6.3 vs. 5.8), ratings for *"some"* showed the opposite pattern (5.3 vs. 5.6). These results suggest that richer discourse context promotes the use of scalar descriptions, specifically

for *"some."* This highlights a key difference in the materials used by GKCT and the current study.

Taken together, these results demonstrate that the effects that we observed in Experiment 2 cannot be attributed to the naturalness hypothesis. When participants were given the stories that accompanied our commands, the use of numbers fillers did not decrease the perceived naturalness of *"some"* as a description of the subset. Moreover, when the stories were removed, making the stimuli more parallel to those used in the GKCT, we replicated some, but not all, of the effects that they observed. Specifically, participants rated *"two"* as a better description of the subset than *"some,"* and the ratings for *"some"* showed a (marginal) decline when numbers were present. However, these patterns were also present for *"all,"* which was rated as worse than *"three"* and received (non-significant) lower ratings in the presence of numbers. Thus we did not observe the critical interaction between Predictability and Quantifier Strength. These results suggest that while the items included in an experiment can sometimes modify naturalness judgments, these effects do not show the pattern that would be necessary to account for the differences in the time-course of processing *"some"* and *"all."* Critically, the differences disappear when richer discourse contexts are provided and thus they cannot explain the patterns that we observed in Experiments 1 & 2 and HS.

Critically, the pattern of judgments is fully consistent with the verbal-encoding hypothesis, which makes only very weak predictions about the naturalness of the descriptions. First, on this hypothesis we assume that participants in scalar-filler conditions will spontaneously encode targets using scalar terms as soon as the sets are presented (*"some"* for the subset and *"all"* for the total set). As they hear the instructions, this pre-encoding allows them to rapidly use the lexical item to directly infer the referent, without generating a scalar implicature. Thus we make

the minimal prediction that both *"some"* and *"all"* will be rated as highly natural in the scalar filler conditions, and they are. Second, under this hypothesis, we assume that in the number filler condition, there are two encodings that are both salient. This will prevent participants from making a direct mapping between the referent and its description, resulting in slower reference resolution when a scalar implicature is necessary. Thus we should expect that both the numbers and scalar terms will receive reasonably high ratings under these circumstances (which they do).

While the verbal-encoding hypothesis makes minimal predictions about naturalness ratings, it makes strong predictions about how participants will describe sets under different testing conditions. Specifically, our hypothesis is grounded in the assumption that variation in the predictability of referring expressions will strongly affect how participants choose to encode the sets. In Experiment 4, we test this assumption by eliciting descriptions of the critical targets. This experiment occurred in two parts. First, participants were presented with the stories and displays, paired with descriptions for total sets and subsets. The predictability of these descriptions was varied by using utterances involved scalar expressions only or a mix of scalar and number words. Next, participants saw new stories and displays and were asked to generate descriptions of their own for target sets. If the verbal-encoding hypothesis is correct, then participants will consistently produce scalar terms when scalar descriptions are highly predictable but will show no clear preference for scalars over numbers when they are not. Critically, this drop in scalar usage should be present not only when the target is the subset (*"some"*) but also when the target is the total set (*"all"*). However, if the naturalness hypothesis is correct (i.e., people strongly prefer to use numbers rather than *"some"* in the number filler condition, but do not have a strong preference for numbers relative to *"all"*), then the use of

number fillers should substantially decrease the production of scalar description for subsets, but not total sets.

5. Experiment 4

5.1. Methods

5.1.1. Participants

Twenty-eight English-speaking undergraduate students at Harvard University and the University of North Carolina at Chapel Hill participated in this study. They received course credit or $10 for their participation.

5.1.2. Procedure

Participants sat in front of a computer display. At the beginning of the study, they were told that they would hear a story describing events about two pairs of characters. After each story, one of the characters would be highlighted with a red box. Their task was to produce an instruction that would allow an imaginary listener to identify that same character in the display. They were then told that they would see several examples of how this could be done. During the familiarization phase, participants saw stories and displays, paired with descriptions for the highlighted character. During the test phase, they were presented with similar stories and displays and were now asked to provide an appropriate written description of their own.

5.1.3. Materials

The familiarization phase borrowed materials from eight stories and displays in the Experiment 1 filler trials. Half of the items highlighted a character with 2-of-4 items (subset) while the other half highlighted one with 3-of-3 items (total set). We can conceive of the four conditions as representing the four cells of 2 x 2 design. The first factor, Predictability, varied the labels used to describe the Target. This was manipulated between subjects. In the scalar

filler condition, participants heard quantities labeled with terms from only the quantifier scale (*"some"* or *"all"*) while in the number filler condition, they heard labels from both the quantifier and number scales (*"some," "all," two,"* and *"three"*). These terms appeared in the same carrier phrase used in Experiment 1 and were written at the bottom of the display (*"Click on the girl/boy that got ____ of the ____"*). The test phase borrowed materials from eight stories and displays in the Experiment 1 critical trials. The second factor, Display Type, varied the target set to be described. Half of the items highlighted the subset while the other half highlighted the total set. This was manipulated within subjects and counterbalanced across two presentation lists.

5.2. Results and Discussion

We focused our analyses on descriptions of the target set that used number words or scalar quantifiers. In the number filler condition, these utterances accounted for 62% of the descriptions for subsets and 65% for total sets. In the scalar filler condition, they accounted for 62% for subsets and 78% for total sets. There were no significant differences across Display Type or Predictability (all p's > .15). All remaining descriptions referred to the target set using only the final noun (e.g., "Click on the girl that has socks"). These were excluded from further analyses.

INSERT FIGURE 5 ABOUT HERE

Our primary dependent measure was the preference to produce a scalar description. For each trial, this value was 1 if participants produced a scalar quantifier (*"some"* or *"all"*) and 0 if they produced a number word (*"two"* or *"three"*). These data were analyzed using a series of logistic mixed-effects models. Subjects and items were included as simultaneous random effects on the intercept only. We found that participants were more likely to generate scalar descriptions in the scalar filler condition than in the number filler condition, leading to a main effect of Predictability (92% vs. 41%; $z = 2.12$, $p < .05$). There was no additional main effect of

or interaction with Display Type (all *p's* > .40). Planned comparisons within Display Type revealed that participants generated more scalar descriptions for the subsets in the scalar filler condition than in the number filler condition (92% vs 34%; z = 2.09, $p < .05$). This same pattern was also found for descriptions of the total sets, although this difference was only marginally significant (90% vs 55%; z = 1.76, $p < .10$). These results suggest that the predictability of set descriptions in the familiarization phrase affected participants' construal of these sets during the test phase. When quantities were labeled with scalar expressions only, participants were far more likely to produce scalar descriptions for both subsets and total sets.

Next we examined whether preferences for scalar descriptions were greater than what would be predicted by chance (in this case 50%). In the scalar filler condition, the preference for the scalar description exceeded chance for both descriptions of subsets (t = 7.07, $p < .001$) and total sets (t = 7.82, $p < .001$). This suggests that exposure to predictable labels led participants to adopt them as the dominant descriptions for these sets. In contrast, in the number filler condition, there were no preferences for either scalar or number descriptions for both subsets and total sets (all *p's* > .40). This suggests that while the number fillers increased production of number words, these descriptions were by no means the dominant response. In fact, participants were just as likely to construe these sets in terms of scalar descriptions.

The findings from Experiment 4 have implications for the interpretation of the eye-tracking results. On the one hand, these findings are problematic for GKCT who attribute prior evidence of delays for *"some"* to a preference to describe the subset as *"two."* The equivocality of participants' utterances in the number filler trials fails to support the notion that number words provide a more salient encoding of small subsets. This strongly suggests that other factors (such as the initial unavailability of scalar implicatures) are responsible for such delays. On the other

hand, these results support the idea that over the course of a few trials, consistent labeling of a quantity as *"some"* can rapidly facilitate the encoding of the referent as a subset. This type of encoding may subsequently enable participants to rapidly restrict the referent of *"some,"* explaining why in some cases, reference restriction occurs as quickly as with *"all."*

6. General Discussion

These experiments explored two questions about why scalar implicatures in the visual world paradigm are sometimes slow (Huang & Snedeker, 2009a, 2011a; Panizza et al., 2009) and sometimes fast (Grodner et al., 2010; Breheny et al., 2013ab). Our focus was on the factors that could explain the discrepancy between GKCT and HS. First, we tested whether rapid scalar implicatures were associated with the prosodic form of the quantifier or the presence of number trials in the task. Experiments 1 and 2 found no evidence that the prosody of *"some"* had any effect on listeners' speed of interpretation. Instead listeners demonstrated a rapid preference for the subset when this quantity was consistently labeled with *"some"* and a delay in interpreting *some* as upper-bounded when descriptions of the subset alternated between *"some"* and *"two."*

Second, we explored two hypotheses about why including number trials might have this effect. GKCT's naturalness hypothesis proposes that the slow-down observed for *"some"* has nothing to do with scalar implicatures but instead reflects the unnaturalness of this description in the presence of numbers. On this hypothesis, we would expect that the inclusion of number trials would result in a sharp decrease in the acceptability of *"some"* but would not have a similar effect on *"all."* In Experiment 3, we looked for such a pattern but did not find it. In rich discourse contexts, like those used in our eye-tracking experiments, *"some"* was perceived as being just as natural when the numbers were present, as it was when the numbers were absent. The naturalness hypothesis also predicts that, when number labels are used, they should become

the dominant descriptions for subsets. But, in Experiment 4, we found that participants were no more likely to label a subset with *"two"* than they were to label it with *"some."*

In contrast, our verbal-encoding hypothesis proposes that, in the absence of the number trials, participants quickly learn to conceptualize the displays in terms of subsets and total sets. This allows them to predict the verbal label that will be used for each set, providing a direct mapping between the lexical form and the referent. When number trials are present, each referent is described in two ways, making verbal encoding a less efficient strategy. Under these conditions, participants must retrieve the meaning of the quantifier to guide reference resolution and then make the scalar implicature before restricting the meaning of *"some."* The results of Experiment 4 lend support to this hypothesis: when participants saw sets predictably labeled with scalar quantifiers during the familiarization trials, they consistently produced these forms. In contrast, when these labels were unpredictable, they were equally likely to produce both quantifiers and numbers. This suggests that no single conceptualization of the stimuli was dominant in this case. Altogether, the current findings are problematic for the naturalness hypothesis but consistent with the verbal-encoding hypothesis.

In the remainder of the General Discussion, we will flesh out the mechanisms involved in the verbal-encoding hypothesis by linking this account to current theories of comprehension and production in psycholinguistics (MacDonald, 1999; Gennari & MacDonald, 2009; Dell & Chang, 2013; Pickering & Garrod, 2013). Next we will discuss how verbal encoding relates to the phenomenon of implicit naming found in prior research (Jescheniak & Levelt, 1994; Meyer, Sleiderink, & Levelt, 1998; Meyer & VanDerMeulen, 2000; Zelinsky & Murphy, 2000; Meyer, Belke, Telling, & Humphreys, 2007). Finally, we will explore whether the verbal-encoding

hypothesis can account for the evidence of rapid scalar implicatures found in other recent studies (Breheny et al., 2013ab).

6.1. Verbal-encoding hypothesis

The current findings suggest that there are two distinct routes by which pragmatic inferences arise (Figure 6). When the prior context is unpredictable, listeners rely on the bottom-up activation of phonological inputs to trigger the relevant lexical entries. If the semantic interpretation of the expression is enough to identify a unique referent in the discourse (e.g.,*"all"* and number words), then reference restriction will be quite rapid. If, however, a pragmatic inference is necessary to do so (e.g., the upper-bound of *"some"*), then reference restriction will be delayed. In contrast, when the context is predictable, listeners will verbally encode the referents using likely descriptions before hearing the critical instruction. These internal, context-specific descriptions are subsequently compared to the external inputs from the speech stream, leading to rapid semantic *and* pragmatic interpretation. Critically, what appears to be instantaneous pragmatic interpretation reflects the listeners' ability to directly access the intended referent, without having to first retrieve the meaning of the quantifier and then make the implicature.

INSERT FIGURE 6 ABOUT HERE

Verbal encoding of this kind bears resemblance to contemporary theories arguing for a tight coupling between language production and comprehension systems. For example, the Production-Distribution-Comprehension framework suggests that the comprehension difficulties associated with particular linguistic structures reflect the likelihood that they are encountered in listeners' input (MacDonald, 1999; Gennari & MacDonald, 2009). Since these statistical profiles are created by speakers' tendency to recruit these structures, this creates an inherent relationship

between comprehension and production across individuals. However, even within an individual, recent accounts have argued that production processes may facilitate comprehension by allowing listeners to anticipate the up-coming input. The P-chain framework suggests that comprehension relies on predictive processes that are implemented through the production system (Dell & Chang, 2013). Similarly, Pickering and Garrod (2013) suggest that predictions during comprehension occur through simulation of utterance production.

Support for these proposals come from striking evidence of early sensitivity to word forms in highly predictable contexts (Visser, Chwilla, & Kolk, 2006; Dikker, Rabagliati, & Pylkkänen, 2009; Dikker, Rabagliati, Farmer, & Pylkkänen, 2010; Dikker & Pylkkänen, 2011, 2013; Kim & Lai, 2012). For example, Dikker and colleagues (2011, 2013) found that pictures with high cloze-probabilities (e.g., banana) can preactivate corresponding lexical entries and word forms (e.g. *"banana"*), leading to a M100 response over the visual cortex when mismatching words are subsequently presented (e.g., *"apple"*). In contrast, this response was not found following pictures with low cloze-probabilities (e.g., grocery), suggesting that the early mismatch effect is specifically driven by the ease of verbally encoding the initial picture. Similar early-emerging P130 effects have been found when reading contextually-supported pseudowords (e.g., *"bake a ceke"*) but not for unsupported words (e.g., *"bake a tont"*) (Kim & Lai, 2012). Altogether, these results are consistent with the current findings and highlight the pervasive use of higher-level information to generate rapid, lower-level lexical predictions.

However, despite the broad consensus that real-time comprehension takes advantage of production-driven predictive processes, there are three ways in which the verbal-encoding hypothesis can be misconstrued. First, one could interpret us as saying that predictable contexts (e.g., ones in which only scalar terms are used) allow participants to develop an experiment-

specific strategy, and thus we should only study scalar implicatures in contexts where the description of each referent varies from trial to trial. This is not our intention. The effects of predictability emerge early in the study and do not grown any stronger over time. They may be strategic, in the sense of being shaped to the task at hand. But they also appear to be deployed rapidly and quite consistently. Thus we prefer to construe these effects as evidence that language processing is dynamic and adaptive. In other words, strategies this good warrant some respect.

Second, our claim that instantaneous implicature only takes place when pre-encoding has occurred could be construed as a claim about modularity. Specifically, one could misinterpret this as a claim that there is a discrete stage of semantic analysis that is independent of pragmatics and must be completed before an implicature can occur. We see no reason to take such a strong position. Incrementality and interactivity appear to be the norm in both language and vision (MacDonald, 1999; Gennari & MacDonald, 2009; Visser et al., 2006; Dikker et al., 2009; Dikker et al., & Pylkkänen, 2010; Dikker & Pylkkänen, 2011, 2013; Kim & Lai, 2012; Huang & Snedeker, 2011b; Huang & Gordon, 2011; Huang & Snedeker, 2013; Dell & Chang, 2013 Pickering & Gerrod, 2013), and there are good reasons to think that the process of calculating a scalar implicature can begin at the scalar trigger (Bergen & Grodner, 2012; Breheny et al., 2006). Our claim is far more modest. We are simply arguing that scalar implicatures are computations that take place in real time, rather than as stored products, and thus accessing the upper-bounded interpretation cannot be instantaneous. Furthermore, when scalar implicatures appear to be timeless, this is because the work was done beforehand. We remain agnostic as to why they take as long as they do—we have consistently found a 600-800ms delay in our visual-world tasks—but we suspect that this difficulty reflects the contextual richness and complexity of these implicatures (Bergen & Grodner, 2012; Breheny et al., 2006; Breheny et al., 2013ab).

Finally, our claim that the lexical meaning of *"some"* is often available before the pragmatic upper bound, in no way implies that we think that pragmatic processes are less important, more fragile, or independent from the processes involved in semantic decoding. We are simply arguing that the path from sound to inference passes through the lexicon, a claim that we hope is uncontroversial.

6.2. Relationship to implicit naming

The current findings also connect to an extensive literature on explicit and implicit naming. Previous research has found that during language production, visual attention is mediated by the selection of possible word forms to describe a referent (Meyer et al., 1998; Meyer & VanDerMeulen, 2000; Griffin, 2001). Pictures with low frequency names generate longer fixation durations compared to those with high frequency names (Meyer et al., 1998; Griffin, 2001). Similarly, pictures with less name agreement elicit longer gaze times compared to those with more name agreement (Griffin, 2001). However, even more striking are effects of language-mediated attention in in tasks that involve no verbal inputs or responses (Jescheniak & Levelt, 1994; Zelinsky & Murphy, 2000; Meyer et al., 2007). For example, Zelinsky and Murphy (2000) found that fixations to the pictures during a memory task are correlated with the length of their names: items with 3-syllable names like *"elephant"* generate longer looking times compared to items with 1-syllable names like *"ball."* Similarly, Meyer and colleagues (2007) found that when asked to verify whether a picture was present in a display (e.g., a bat-animal), participants often generate eye-movements to an item with a shared name (e.g., a baseball bat). Additional evidence of implicit naming in 18-month-old infants provides additional support that access to linguistic labels through verbal encoding is fairly automatic and is unlikely to be strategically driven (Mani & Plunkett, 2010).

However, a recent study does suggest that explicit and implicit naming differ along one important dimension, specifically whether they activate phonological representations. Pontillo and colleagues (2013) compared phonological cohort effects in two word recognition tasks. Similar to a standard visual-world paradigm, participants in the explicit naming condition were asked for one of two targets (e.g., *"cow"* or *"soda"*) while eye-movements were measured to a competitor that could adopt either a dominant (e.g., *"couch"*) or subordinate label (e.g., *"sofa"*). Both instructions led to comparable cohort effects, suggesting that explicit naming activates phonological associates of both dominant (e.g., *"cow"* → *"couch"*) and subordinate labels (e.g., *"soda"* → *"sofa"*). In contrast, in the implicit naming condition, a visual mask occluded the competitor shortly after the display appeared. This manipulation encouraged participants to actively encode the picture names with verbal labels prior to the instruction. Critically, following the onset of the instruction, cohort effects were found for targets associated with the dominant label but not the subordinate label. This suggests that implicit naming fails to activate the full range of phonological representations seen in a standard visual-world paradigm.

Nevertheless, these results do not address whether implicit naming activates representations for other aspects of linguistic interpretation. Critical to the current hypothesis is whether verbal encoding increases the availability of lexical representations. Recall that the verbal-encoding hypothesis suggests that the predictability of hearing subsets described as *"some"* increases the likelihood that listeners will spontaneously encode these referents in these terms. This direct, pre-activation of lexical representations facilitates interpretation in situations that require a semantically-mediated scalar implicature. The current findings reveal that verbal encoding specifically benefits the interpretation of *"some,"* providing strong support for this account.

After all, if verbal encoding also pre-activated phonological representations, then the rapidity of interpretation should have improved following *"all"* and number words as well.

6.3. Accounting for other evidence of rapid implicatures

The current findings contribute to a growing body of work demonstrating that scalar implicatures unfold over time. Sentence judgments take longer when they include the implicature compared to when they do not (Rips, 1975; Noveck & Posada, 2003; Bott & Noveck, 2004; De Neys & Schaeken, 2007). Similarly, inferred access to the complement set (*"the rest"*) via a scalar implicature emerges only when an extended delay follows the scalar trigger (Breheny et al., 2006; Bergen & Grodner, 2012; Nieuwland et al., 2010; Hartshorne & Snedeker, under review). Moreover, the current findings demonstrate that even scalar implicatures that appear instantaneous take time to generate. Far from immediate, these early pragmatic interpretations simply reflect processes that were undertaken before the start of the utterance.

The current study focused on resolving differences between findings by HS and GKCT. Critically, in this final section, we examine whether the verbal-encoding hypothesis also explain other evidence of rapid scalar implicatures. We focus on two recent visual-world experiments by Breheny and colleagues (2013ab). In the first study (Breheny et al., 2013a), participants were familiarized with events involving quantities transferred into two bowls. During the test phase, participants heard sentences like (7) while their eye-movements were measured to a display featuring half of the limes in one bowl (subset) and all of the oranges in another bowl (total set).

(7) The man has poured <u>some of the water</u> with limes into the bowl on tray A and all of the water with oranges into the bowl on tray B.

Following the offset of *"some,"* participants rapidly shifted looks to the subset, suggesting that an early-emerging scalar implicature was generated to rule out the total set. Similar to GKCT,

reference resolution in these trials was comparable to the *"all"* trials and was substantially faster than control trials asking for *"some"* in the presence of two subsets.

However, a closer examination suggests that features of this study were highly conducive to verbal encoding. First, each quantity was described in a highly uniform manner in the critical sentences: subsets as *"some"* and total sets as *"all."* In fact, across all trials, labels for the referent could be perfectly predicted based on the quantity of the target set. Second, sentence (7) illustrates that sets were exhaustively labeled within each trial, providing direct contrast between quantities. This may have promoted the salience of set labels as well as increased the benefits of adopting these labels for identifying multiple referents. Finally, since familiarization events unfolded slowly over time (each lasting approximately 25 seconds), participants were provided ample opportunity to generate appropriate labels for referents. Critically, eye-movements data provide additional support that participants actively engaged in verbal encoding. Following the offset of the quantifier, switches off the target in the critical *"some"* trials were higher than the *"all"* trials and comparable to those in the control *"some"* trials. Consistent with prior findings (Huang & Snedeker, 2009a; 2011a), these results suggest that listeners entertain the total set as a possible referent of *"some"* until 600-700ms after the *offset* of the quantifier.

A second study examined the generation of particularized implicatures involving non-conventional scales (Breheny et al., 2013b). Participants heard sentences like (8) while their eye-movements were measured to a display featuring a fork placed in box A (1-object, subset) and a fork and a spoon placed in box B (2-object, superset).

(8) The woman put a <u>fork into box</u> A and a fork and a spoon into box B.

Similar to *"some,"* the onset of *"fork"* introduces a temporary ambiguity between the subset and superset. Nevertheless, immediately after the onset of *"box,"* participants made predictive eye-

movements to the subset containing a fork and nothing else. This suggests rapid calculation of a particularized implicature. However, the earliness of this inference again may have reflected properties of the study that promoted verbal encoding. Like Breheny and colleagues (2013a), familiarization events unfolded over an extended period of time, and instructions contrasted labeled for both sets within a single trial. Critically, subsets were always described in a uniform manner across all trials (e.g., singular NPs like *"a fork"*), making relevant labels for this target highly predictable. In contrast, recent work adopting variable labels have found consistent delays for generating both generalized and particularized implicatures (Huang, 2014).

6.4. Conclusions

Understanding the processing demands of language comprehension requires an awareness of the procedures that contribute to interpretations along different time scales. This includes not only the moment-to-moment changes that follow a linguistic stimulus but also the expectations that listeners generate prior to this point. By examining the test case of scalar implicatures, the current study suggests the presence of dual mechanisms for interpretation: 1) a rapid mapping from the form directly to the referent facilitated by verbal encoding; 2) a slower process of calculating the upper-bound from the lexical meaning via a pragmatic inference. Understanding the dynamic interplay between these processes will require further experimental research to tease apart these procedures as well as computational models to formalize these concepts.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language,* 38, 419-439.

Altmann, G., & Kamide, Y. (2004). Now you see it, now you don't: Mediating the mapping between language and the visual world. In: J. Henderson & F. Ferreira (Eds.). *The Interface of Language, Vision, and Action*, 347-386. New York: Psychology Press.

Bates, D. M. (2007). Linear mixed model implementation in lme4. *Manuscript, University of Wisconsin - Madison*, January 2007.

Beckman, M. E., & Hirschberg, J. (1994). *The ToBI annotation conventions.* Columbus, OH: Ohio State University.

Bergen, L., & Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38, 1450.

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

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

Breheny, R., Ferguson, H. J., & Katsos, N. (2013a). Investigating the timecourse of accessing conversational implicatures during incremental sentence interpretation. *Language and Cognitive Processes*, 28, 443-467.

Breheny, R., Ferguson, H. J., & Katsos, N. (2013b). Taking the epistemic step: toward a model of on-line access to conversational implicatures. *Cognition*, 126, 423-440.

Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatic interface. In A. Belletti (Ed.), *Belletti structures and beyond*. Oxford: Oxford University Press.

Degen, J., & Tanenhaus, M. (in press). Processing scalar implicature: A constraint-based approach. To appear in *Cognitive Science*.

Dell, G. S., & Chang, F. (2014). The P-chain: relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369, 20120394.

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology,* 54, 128-133.

Dikker, S., Rabagliati, H., & Pylkkänen, L. (2009). Sensitivity to syntax in visual cortex. *Cognition*, 110, 293-321.

Dikker, S., Rabagliati, H., Farmer, T. A., & Pylkkänen, L. (2010). Early occipital sensitivity to syntactic category is based on form typicality. *Psychological Science*, 21, 629-634.

Dikker, S., & Pylkkanen, L. (2011). Before the N400: Effects of lexical–semantic violations in visual cortex. *Brain and Language*, 118, 23-28.

Dikker, S., & Pylkkänen, L. (2013). Predicting language: MEG evidence for lexical preactivation. *Brain and Language*, 127, 55-64.

Feeney, A., Scafton, S., Duckworth, A., & Handley, S. J. (2004). The story of some: Everyday pragmatic inferences by children and adults. *Canadian Journal of Experimental Psychology*, 58, 121-132.

Gennari, S. P., & MacDonald, M. C. (2009). Linking production and comprehension processes: The case of relative clauses. *Cognition*, 111, 1-23.

Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition*, 82, B1-B14.

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

Gadzar, G. (1979). *Pragmatics: Implicature, presupposition and logical form*. New York: Academic Press.

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

Hartshorne, J. & Snedeker, J. (under review). The speed of inference: Evidence against rapid use of context in calculation of scalar implicatures.

Huang, Y. & Snedeker, J. (2009a). On-line interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology*, 58, 376-415.

Huang, Y. & Snedeker, J. (2009b). Some questions are still unresolved: Prosody, predictability, and speed of scalar implicatures. Paper presented at the 2009 Experimental Pragmatics Conference. Lyon, France.

Huang, Y., Hahn, N., & Snedeker, J. (2010). Some inferences still take time: Prosody, predictability, and the speed of scalar implicatures. Poster presented at the 23rd annual CUNY conference on Human Sentence Processing. New York, NY.

Huang, Y. & Snedeker, J. (2011a). 'Logic & Conversation' revisited: Evidence for a division between semantic and pragmatic content in real time language comprehension. *Language and Cognitive Processes*, 26, 1161-1172.

Huang, Y. & Snedeker, J. (2011b). Cascading activation across levels of representation in children's lexical processing. *Journal of Child Language*, 38, 644-661.

Huang, Y. & Snedeker, J. (2013). The use of referential context in children's on-line interpretation of scalar adjectives. *Developmental Psychology*, 49, 1090-1102.

Huang, Y. (2014). Pragmatic inferencing across scales: Linguistic and extra-linguistic effects. Poster presented at the 27th annual CUNY conference on Human Sentence Processing. Columbus, OH.

Huang, Y. & Gordon, P. (2011). Distinguishing the time-course of lexical and discourse processes through context, co-reference, and quantified expressions. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 37, 966-978.

Horn, L. (1972). *On the semantic properties of the logical operators in English*. Doctoral dissertation, UCLA, Los Angeles, CA. Distributed by IULC, Indiana University, Bloomington, IN.

Horn, L. (1989). *A natural history of negation*. Chicago, IL: University of Chicago Press.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 20, 824–843.

Kim, A., & Lai, V. (2012). Rapid interactions between lexical semantic and word form analysis during word recognition in context: evidence from ERPs. *Journal of Cognitive Neuroscience*, 24, 1104-1112.

Ladusaw, W. A. (1994). Thetic and categorical, stage and individual, weak and strong. In M. Harvey and L. Santelmann (eds.) *Proceedings from Semantics and Linguistic Theory IV*. Cornell University, Department of Modern Languages and Linguistics.

Levinson, S. (2000). *Presumptive meanings*. Cambridge, MA: MIT Press.

MacDonald, M. C. (1999). Distributional information in language comprehension, production, and acquisition: Three puzzles and amoral. In B. MacWhinney (Ed.), *The emergence of language*. Mahweh, NJ: Lawrence Erlbaum Associates Publishers

Mani, N. & Plunkett, K. (2010). In the infant's mind's ear: Evidence for implicit naming in 18-month-olds. *Psychological Science*, 21, 908-913.

Matin, E., Shao, K.C., & Boff, K.R. (1993) Saccadic overhead: Information processing time with and without saccades. *Perception & Psychophysics*, 53, 372-380.

Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, 66, B25-B33.

Meyer, A. S., & van der Meulen, F. F. (2000). Phonological priming effects on speech onset latencies and viewing times in object naming. *Psychonomic Bulletin & Review*, 7, 314-319.

Meyer, A. S., Belke, E., Telling, A., & Humphreys, G. W. (2007). Early activation of object names in visual search. *Psychonomic Bulletin and Review*, 14, 710-716.

Nieuwland, M., Ditman, T., & Kuperberg, G. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language*, 63, 324-346.

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

Panizza, D., Chierchia, G., Huang, Y., & Snedeker, J. (April, 2009). Relevance of polarity for the on line interpretation of numerals and determiners. Paper presented at the 19th annual Semantics and Linguistic Theory (SALT) conference. Columbus, OH.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36, 329-347.

Pontillo, D., Salverda, A. P., & Tanenhaus, M. K. (2013). Implicit naming in the visual world paradigm. Poster presented at the 26th annual CUNY conference on Human Sentence Processing. Columbia, SC.

Postal, P. (1964). Limitations of Phrase Structure Description. In J. K. a. J. Fodor (Ed.), *Readings in the Philosophy of Language*. Englewood Cliffs, NJ: Prentice-Hall.

Rips, L. J. (1975). Quantification and semantic memory. *Cognitive Psychology*, 7, 307-340.

Sedivy, J., Tanenhaus, M., Chambers, C., & Carlson, G. (1999). Achieving incremental semantic interpretations through contextual representation. *Cognition*, 71, 109-147.

Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632.

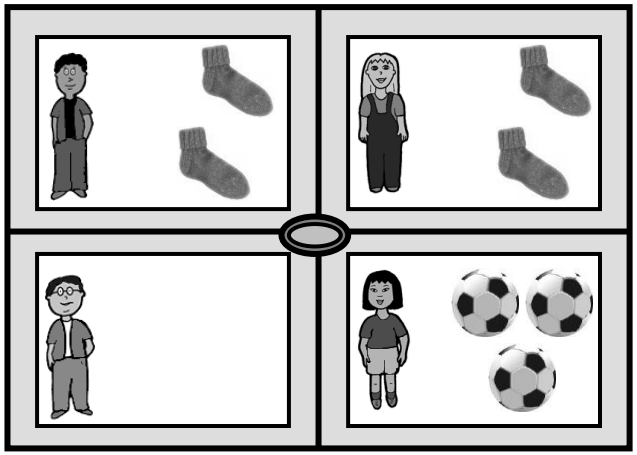
Trueswell, J. & Tanenhaus, M. (1994). Toward a lexicalist framework of constraint-based syntactic ambiguity resolution. In C. Clifton & L. Frazier (Eds.), *Perspectives on sentence processing*. Hillsdale, NJ: Lawrence Erlbaum.

Vissers, C. T. W., Chwilla, D. J., & Kolk, H. H. (2006). Monitoring in language perception: the effect of misspellings of words in highly constrained sentences. *Brain research*, 1106, 150-163.

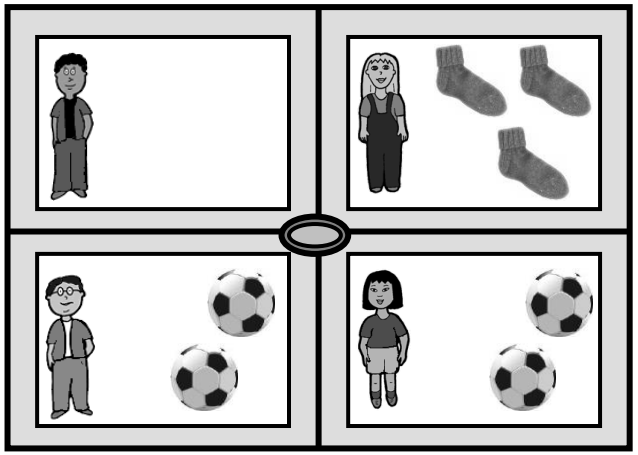
Zelinsky, G. & Murphy, G. (2000). Synchronizing visual and language processing: An effect of object name length on eye movements. *Psychological Science*, 11, 125-131.

Figure 1. In Experiment 1, example of 2-referent displays for (A) *"summa"* trials and (B) *"alla"* trials and 1-referent displays for (C) *"summa"* trials and (D) *"nunna"* trials. Participants here were instructed to *"Click on the girl that got ____ of the socks."* The girl with socks was the Target while the girl with soccer balls was the Distractor.

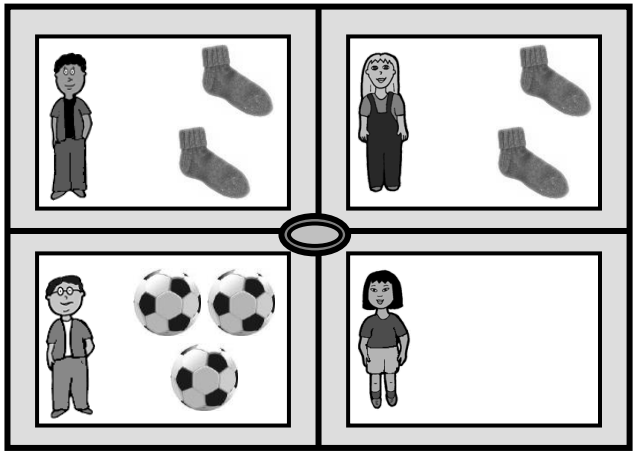
(A)



(B)



(C)



(D)

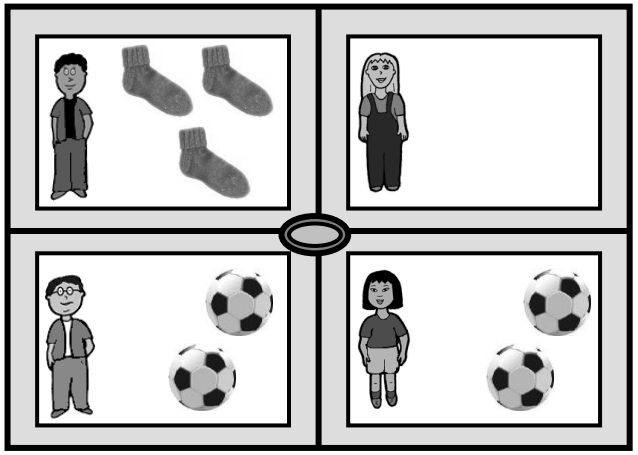


Figure 2. In Experiment 1, the time-course of Target looks in the (A) 1-referent trials and (B) 2-referent trials.

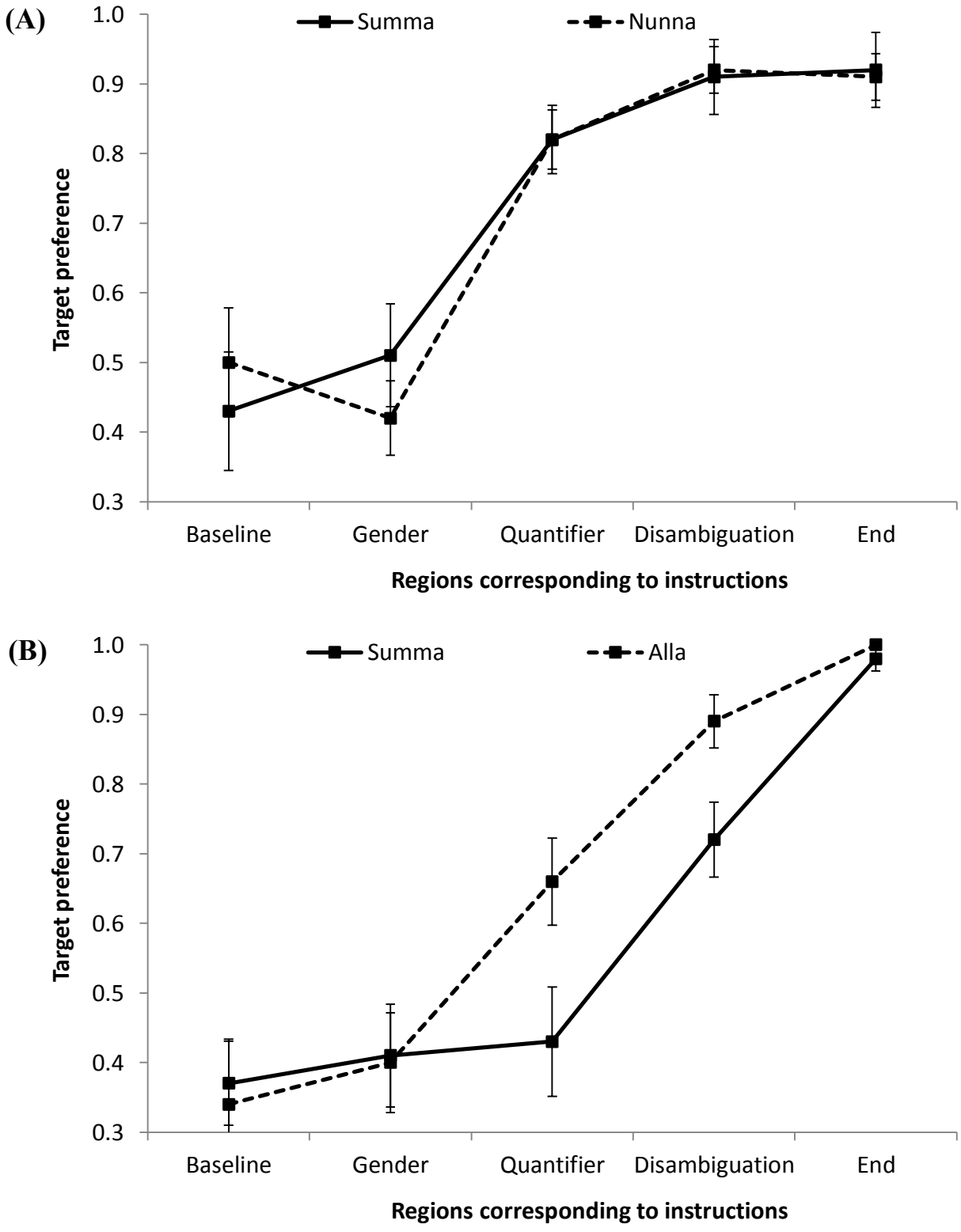


Figure 3. In Experiment 2, the time-course of Target looks in the (A) number filler condition when verbal labels were variable and (B) scalar filler condition when verbal labels were uniform.

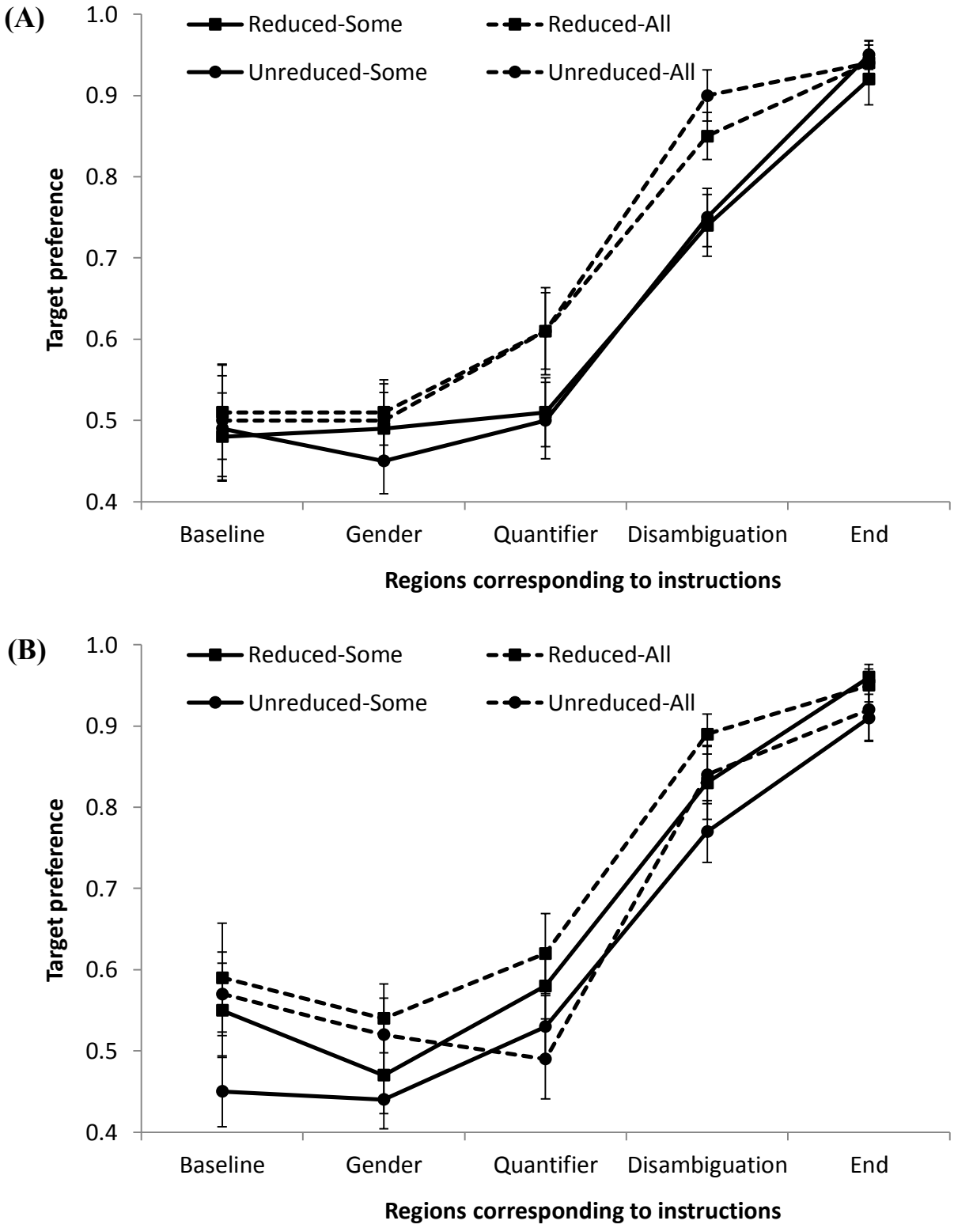


Figure 4. In Experiment 3, the mean naturalness ratings of utterances describing sets using *"two"*, *"some"*, *"three"*, and *"all"* in the (A) trials presented without the story and (B) trials with the story, parallel to those in Experiments 1 and 2.

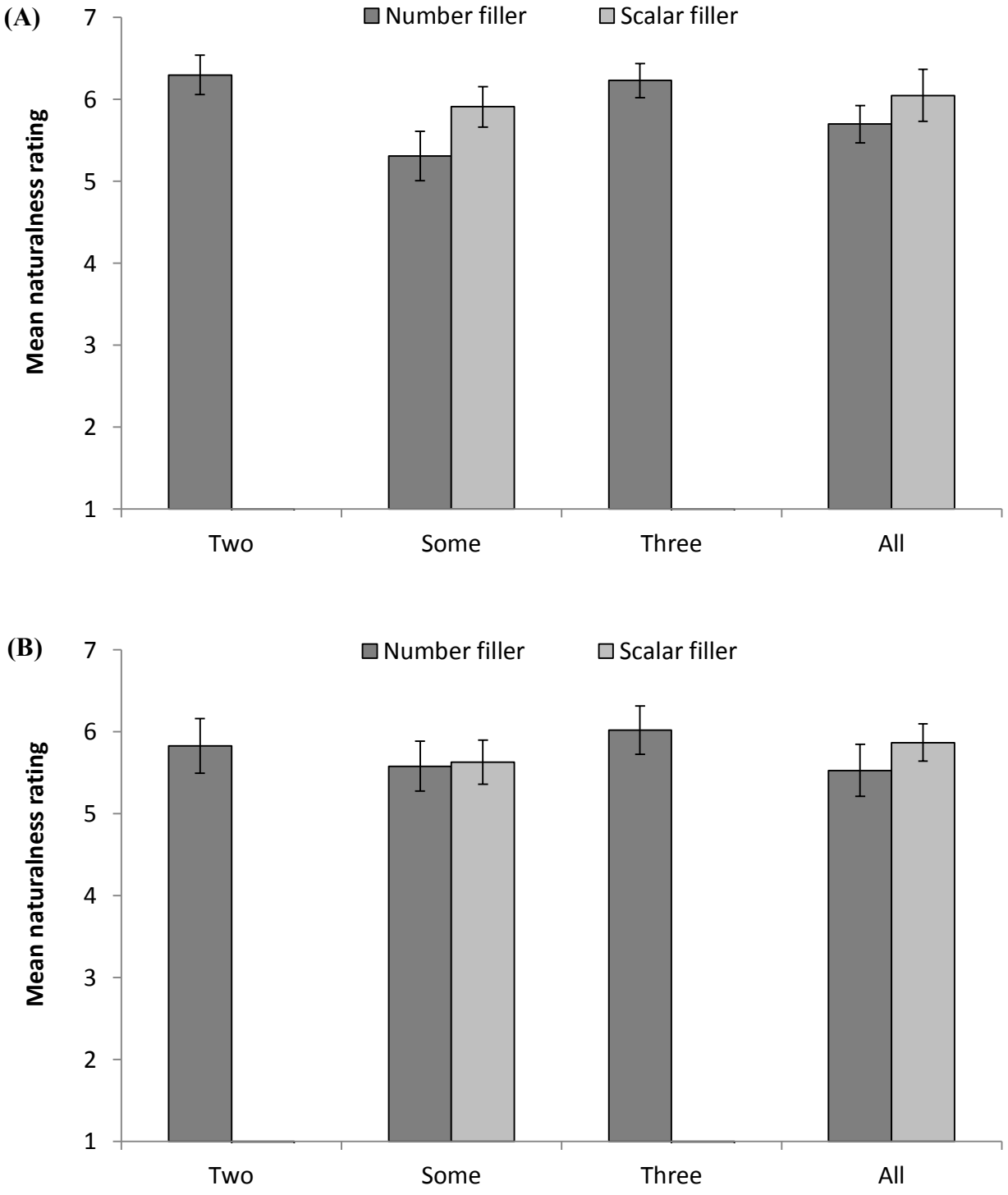


Figure 5. In Experiment 4, descriptions produced for subsets and total in the (A) number filler condition when familiarized verbal labels were variable and (B) scalar filler condition when familiarized verbal labels were uniform.

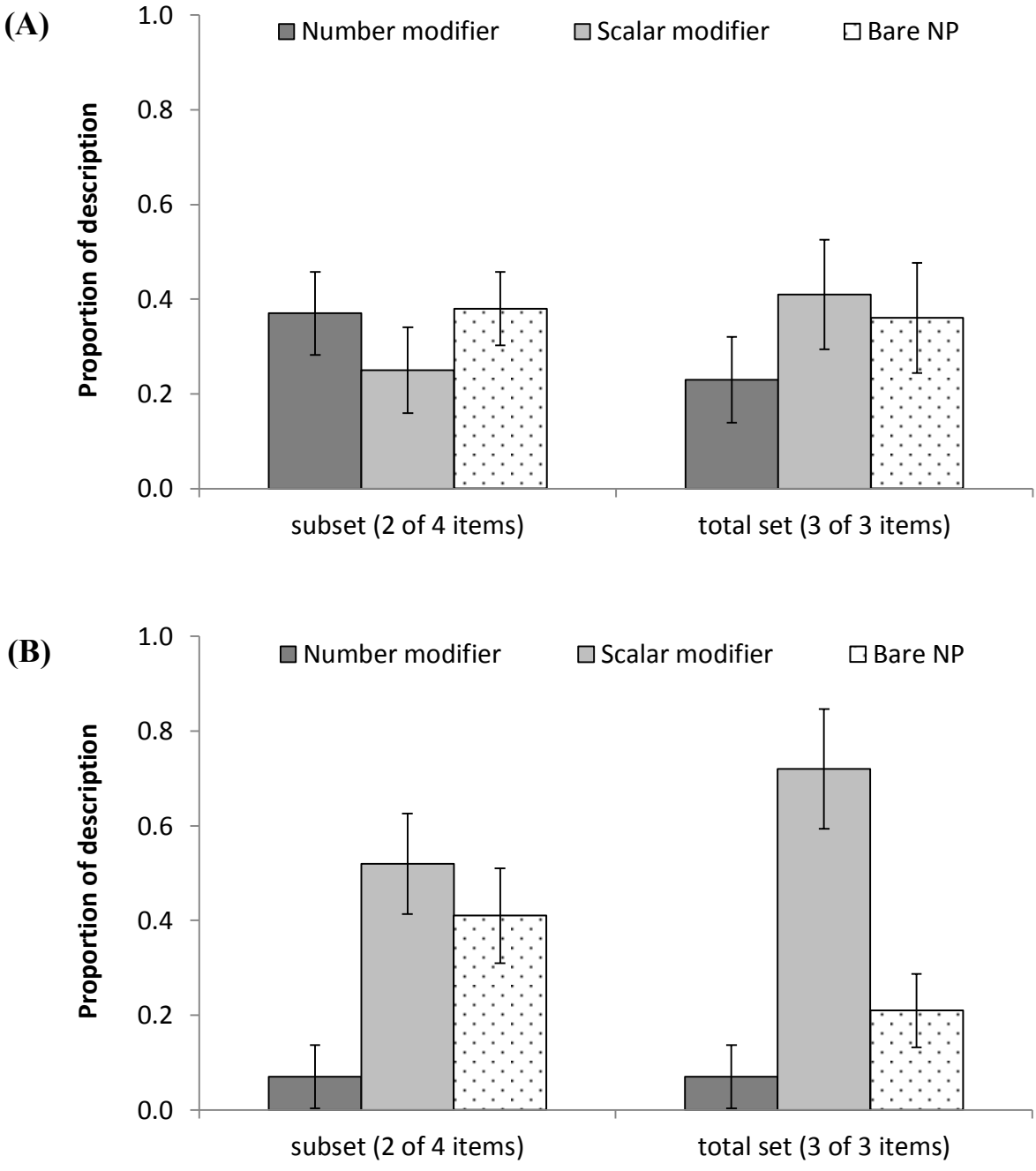
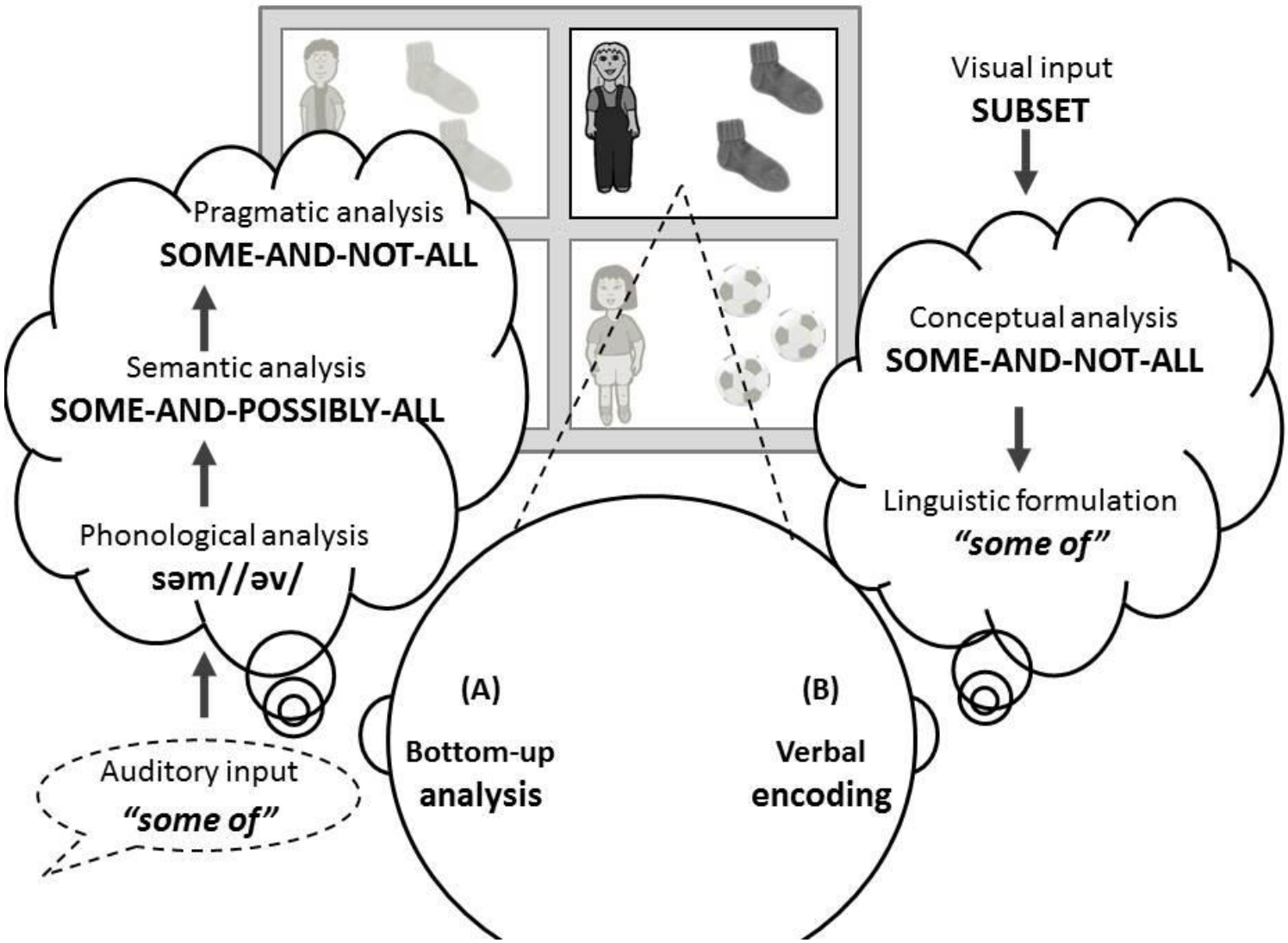


Figure 6. Two pathways by which scalar inferences could arise: (A) Bottom-up analysis driven by the acoustic input and (B) Verbal encoding driven by implicit naming of referents in the display.



Appendix A. Stimuli items for Experiment 1.

Type	Item	Sentence	Distractor
Critical trials	1	Click on the girl that got _____ the pills	pillows
	2	Click on the girl that got _____ the watermelons	waffles
	3	Click on the boy that got _____ the peas	pizzas
	4	Click on the boy that got _____ the robes	roses
	5	Click on the girl that got _____ the sandals	sandwiches
	6	Click on the girl that got _____ the rats	rabbits
	7	Click on the boy that got _____ the mushrooms	muffins
	8	Click on the boy that got _____ the bees	beetles
Filler trials	1	Click on the boy that didn't get anything (paints)	papers
	2	Click on the girl that didn't get anything (socks)	soccer balls
	3	Click on the boy that got two of the cards	cars
	4	Click on the girl that got two of the dogs	dolls
	5	Click on the girl that got two of the turtles	turkeys
	6	Click on the boy that got three of the matches	maps
	7	Click on the boy that got two of the baskets	bats
	8	Click on the girl that got three of the seals	seagulls

Appendix B. Stimuli items for Experiment 2.

Type	Item	Target	Distractor	Type	Item	Target	Distractor
Critical trials	1	pills	pillows	Filler trial manipulation	1	markers	marbles
	2	turtles	turkeys		2	ladders	laptops
	3	sandals	sandwiches		3	bows	boas
	4	papers	paints		4	lemons	lettuce
	5	rats	rabbits		5	flags	flashlights
	6	matches	maps		6	beets	beans
	7	cards	cars		7	berries	bears
	8	watermelons	waffles		8	canes	capes
	9	mushrooms	muffins		9	apples	cookies
	10	dolls	dogs		10	kites	footballs
	11	seals	seagulls		11	medals	trophies
	12	peas	pizzas		12	rings	tiaras
	13	bees	beetles		13	hotdogs	pies
	14	baskets	bats		14	pipes	combs
	15	socks	soccer balls		15	carpets	lamps
	16	robes	roses		16	racquets	bicycles
Filler trial (other gender - empty set)	1	ladybugs	leaves	Filler trial (other gender - subset)	1	tables	couches
	2	bananas	carrots		2	cats	fishes
	3	gloves	skates		3	compasses	rulers
	4	bells	whistles		4	hammers	screwdrivers
	5	purses	wallets		5	cameras	phones
	6	candy	cakes		6	toothpaste	toothbrushes
	7	bottles	cans		7	knives	spoons
	8	plates	cups		8	arrows	bows
	9	candles	light bulbs		9	shoes	balls
	10	coins	keys		10	suitcases	Back bags
	11	peppers	pickles		11	ties	cufflinks
	12	pans	ladles		12	sunglasses	umbrellas
	13	beakers	books		13	toasters	blenders
	14	helmets	jerseys		14	speakers	headphones
	15	belts	watches		15	CDs	ipods
	16	drums	guitars		16	envelopes	notebooks