

A within-subjects comparison of the acquisition of quantity-related inferences

Alicia Parrish & Ailís Cournane*

Abstract. This study directly compares quantity inferences from scalar implicatures (‘Some of the ducks are black’) and uniqueness presuppositions in definites (‘the duck is black’) to exhaustivity inferences in English *it*-clefts (‘It’s the duck that’s black’) for which the theoretical literature disagrees on the source of inference – pragmatic (like scalar implicatures), or semantic (like presuppositions). We investigate whether within-subjects correlations in acquisition can inform us about the source of exhaustivity inferences. Assuming comprehension is achieved once the necessary basis for meaning is acquired, *it*-clefts should pattern with presupposition judgments if computing a presupposition is involved and should pattern with scalar implicature judgments if computing an implicature is involved. We conduct three experiments to test how closely *it*-cleft judgments pattern with other quantity-related inferences, keeping materials maximally similar. The first two experiments test adult participants using a Truth Value Judgment Task and then a 3-point Rating Task; we find that adults’ response patterns to under-informative uses of these constructions differ both across individuals and across inference types, with the Rating Task giving more informative results. In the third experiment, we use the 3-point Rating Task with 4-, 5-, and 6- year olds to characterize response patterns across the three inference types for each individual subject. We find that the individual response patterns children exhibit are consistent with the theory that *it*-cleft exhaustivity shares an underlying cognitive source with the computation of presupposition inferences, but not with scalar implicature inferences.

Keywords. clefts; exhaustivity; acquisition; pragmatics; semantics; experimental

1. Introduction. As part of normal language development, children must master the diverse landscape of inferences that people make in discourse. English sentences with *it*-clefts, such as in (1), carry one such relevant inference, called an ‘exhaustivity inference’. That is, the sentence in (1) gives the interpretation that there is nothing other than the duck that is black, making ‘the duck’ an exhaustive listing of salient black things.

(1) It’s the duck that’s black.

Explanations for the source of exhaustivity inferences in *it*-cleft structures vary, and there is a lack of consensus on where this inference is derived from. Relevant theoretical works tend to fall into one of two camps, placing the source for exhaustivity inferences in the same group as either presuppositions or implicatures. We broadly refer to these two options as ‘semantic’ and ‘pragmatic’ approaches to *it*-cleft exhaustivity¹. Our goal is to characterize the mechanism

*We thank Lucas Champollion for early help developing this idea. Sudha Arunachalam and Philippe Schlenker both provided valuable and detailed feedback. Additional helpful discussion about this project came from NYU’s Child Language Lab and from discussions with members of NYU’s semantics reading group which included Masha Esipova, Alex Warstadt, Anna Alsop, Ioana Grosu, and Paloma Jeretič. We also sincerely thank Daniella Presti, Chiara Repetti-Ludlow, Vishal Arvindam, and Sasha Frangulov for help in data acquisition. Authors: Alicia Parrish, New York University (alicia.v.parrish@nyu.edu) & Ailís Cournane, New York University (cournane@nyu.com).

¹There is another important approach to *it*-cleft exhaustivity that accounts for it as a ‘homogeneity’ inference. We address this briefly in the introduction and return to it in discussion, but do not directly test this possibility.

underlying this exhaustivity inference by directly contrasting it with inferences derived from scalar implicatures and uniqueness presuppositions.

Past studies on the acquisition of exhaustivity inferences have revealed mixed results. Cremers et al. (2016) tested how 5-year-olds understand the ambiguity of weakly exhaustive inferences in questions (those embedded under “know”). They found that although children at this age understand the exhaustive interpretation, they gave responses that reflect a strongly exhaustive reading. They concluded that, at least for children, exhaustivity in questions is derived in a way that is potentially related to scalar implicature judgments for children, and not necessarily through the adult-like application of an exhaustification operator (the same type of operator theorized to be at play for *it*-cleft exhaustivity, among other inferences).

In a study that directly investigated children’s understanding of *it*-clefts, Tieu & Križ (2017) found that children consistently accepted non-exhaustive interpretations², indicating that they may fail to generate or integrate the relevant exhaustivity inference. The children in their study also accepted sentences with presupposition failures of plural definites, though, while adults consistently rejected these items. They concluded that their results are in line with semantic accounts of exhaustivity, specifically relating the inference to children’s understanding of homogeneity requirements. Given the inconclusive findings in the few studies that look at the acquisition of exhaustivity, we directly test whether one instance of exhaustivity inferences, English *it*-clefts, patterns more closely with presuppositions or implicatures in children.

1.1 IT-CLEFT EXHAUSTIVITY AS SEMANTIC. Büring & Križ (2013) attribute the exhaustivity inference in *it*-clefts to semantic content and argue that it is a presupposition of *it*-clefts. In (2), the exhaustivity trigger is embedded under a belief context. On a semantic account, because the exhaustive interpretation associated with ‘only’ is at-issue, the speaker can felicitously deny having known something about that exhaustivity, but the same is not the case for the not-at-issue exhaustivity in (2-b).

- (2) a. I knew that the duck was black, but I didn’t know that only the duck was black.
b. # I knew that the duck was black, but I didn’t know it was the duck that’s black.

Križ (2016) offers a somewhat updated version of this approach, framing the *it*-clefts as “identity statements between two individuals.” On this analysis, they also expect *it*-clefts and definite descriptions to pattern together. While we do not directly test this inference here, this possibility has been addressed by Aravind et al. (2018).

Similarly, Velleman et al. (2012) consider exhaustivity as part of the not-at-issue content in *it*-clefts. On their view, *it*-clefts contrast with only-constructions as ‘only’ presupposes a minimal answer and asserts a maximal one, while *it*-clefts presuppose a maximal answer and assert a minimal one. Though there are slightly different predictions made by these theories, we take a broad approach here, namely, if the underlying cognitive or linguistic mechanism is related to presupposition-based inferences, we expect *it*-clefts or definite descriptions to pattern together with other presuppositions in a measurable way.

Destruel et al. (2015) provide experimental evidence for a system that is at least compatible with that of Velleman et al. (2012). Focusing on the question of what semantic content is at-issue versus not-at-issue, they found that the most acceptable answer to a violation of the

²However, it’s notable that in their study adults also accepted non-exhaustive readings in *it*-clefts about half the time, though still at rates lower than children.

exhaustivity inference in *it*-clefts is to respond with “yes, but...” (or “yes, and...”) over just a straight response of “yes” or “no”, and claim that the at-issue content is accepted when one responds “yes”, but that the exhaustivity, which is not-at-issue, still needs to be addressed.

Building off of the idea that *it*-cleft exhaustivity is literally a type of presupposition, we take as a comparison a sentence of the form in (3), where the form is maximally similar to the *it*-cleft type sentence shown in (1) and the meaning of the inference is similarly related a quantity-based restriction. The relevant presupposition of (3) is that there is a single salient duck, and it is black. Thus any context in which there are two black ducks would be incongruent with this presupposition.

(3) The duck is black.

Past studies that have investigated judgments of presupposition violations have found that even adults are not as consistent in their judgments as one might expect. Cremers et al. (2017) used probability judgment tasks and found that adult participants sometimes treat sentences with a presupposition failure as true. Similarly, on a picture-choice variant of the covered-box task, Bill (2015) found that adults accommodate presupposition violations far more readily than children and suggested that this difference between children’s and adults’ acceptance of sentences with presupposition violations is due to accommodation being an extra step in judging the truth of a sentence, and that children in the 4 to 7 age range are not consistently able to compute this extra step in order to accommodate the violation. Importantly, children do not fail to show adult-like judgments in presuppositions due to a lack of ability to reason about the common ground. Aravind (2018) showed that even four-year-olds can use linguistic cues to reason about an informed vs. uninformed observer. The crucial issue for presuppositions, then, is whether children are able to understand presuppositional content as informative. Aravind (2018) suggests that this is what may be difficult, at least for 4-year-olds.

1.2 IT-CLEFT EXHAUSTIVITY AS PRAGMATIC. Contrary to proposals that account for *it*-cleft exhaustivity with a semantic source, other researchers have proposed that it is a kind of pragmatic implicature, and thus it is computed in addition to the semantic content of the utterance. The implicature view is driven by the observation that actual inferences derived from *it*-cleft exhaustivity are often weaker than what would be predicted based on a purely semantic, presupposition-based account. Destruel (2013), for example, found cases where the exhaustivity inference in French *c’est*-clefts (which carry the same exhaustive interpretation as the English cleft structures (Destruel & Donaldson 2017)) is cancellable, and thus incompatible with its status as a presupposition. The alternative proposal put forth is that the exhaustivity inference in *it*-clefts, unlike exclusive items like “only”, derives from an implicature, where the inference arises from Gricean reasoning (Grice 1975).

To test the degree to which *it*-clefts and “only” diverge, DeVeugh-Geiss et al. (2015) conducted two experiments using acceptability judgment ratings. They tested sentences that contradict the at-issue or not-at-issue content of clefts compared to exclusives (such as “only”), and of exclusives compared to pseudo-clefts. The results overall showed that a contradiction of not-at-issue content in *it*-clefts gets a higher acceptability rating than a contradiction of the not-at-issue content of exclusives, supporting a distinction between these two ostensibly similar inferences. In their conclusion, they explicitly state that they consider the exhaustivity inference in *it*-clefts to be a kind of scalar implicature, and thus attribute it to a purely pragmatic

source. DeVeugh-Geiss et al. (2017) offer an updated account of these findings, though their results are again somewhat inconclusive. They found that neither *it*-clefts nor definite pseudoclefts consistently elicited purely exhaustive interpretations, which is surprising given the regularity of assuming exhaustive interpretations for these constructions in the theoretical literature. Rather, interpretations of what should have been infelicitous uses were also reported to be acceptable some of the time. DeVeugh-Geiss et al. (2017) concluded that there must be both presuppositions and scalar implicatures at play in *it*-clefts and pseudoclefts.

Building off of the most standardly used sentence frame for generating this kind of scalar implicature, we use sentences of the form in (4), which is again kept maximally similar to the *it*-cleft example in (1) in both form and meaning. (4) gives rise to a scalar implicature, such that it is only felicitous when some, but not all, of the ducks are black.

(4) Some of the ducks are black.

Many studies have focused on children's acquisition of scalar implicature judgments using this contrast. Though using "some" when "all" is more informative is logically true, it is underinformative and adults tend to reject these sentences while children accept them at greater rates (Smith 1980; Noveck 2001; i.a.). There is some consensus that children struggle on this task because they either do not know the lexical scale-mate to "some" (Barner et al. 2011), or they fail to access the alternatives in real time (Skordos & Papafragou 2016). "Awareness of the task" also affects whether children behave in an adult-like way, though this may be orthogonal to their understanding of the implicature (Papafragou & Musolino 2003). Both providing the child with examples of the relevant alternatives (Chierchia et al. 2001) and providing substantial contextual support (Foppolo et al. 2012) makes children's judgments more adult-like, though. Crucially, part of the issue from task effects may be due to the use of a truth value judgment task, as this task not only artificially restricts judgments, but also collapses cases that are actually false with those that are just infelicitous. Katsos & Bishop (2011) have shown that using a 3-point scale can be more informative when dealing with judgments that are not actually false, but somewhere in the middle of the scale.

1.3 THE PRESENT STUDY. We test the two different accounts of *it*-cleft exhaustivity detailed in this section. In order to do this, we keep items maximally similar in form. One clear concern with the comparisons being made in this study is that *it*-clefts in English differ from the presupposition and scalar implicature examples that we use in that they are always bi-clausal, and therefore have a more complicated syntax. On the surface, this difference may predict that development of adult-like inferences in *it*-clefts would require additional processing costs, and therefore cause that development to be delayed relative to similarly difficult mono-clausal structures. While this confound is unavoidable given the contrasts we're testing, 3-year-old children already begun productively producing sentences with embedded CP structures, and they show competency in understanding the relationship between verbs and their complements by age four (Diessel & Tomasello 2001; de Villiers & Roeper 2016). Given these findings, the bi-clausal structure in clefts is not itself a barrier to children's ability to comprehend them.

2. Experiments 1 & 2: Adult norming. We conduct two preliminary experiments to assess the concern about task choice and to characterize the adult responses on our experimental items. We do not yet know the *within-subjects* patterning of SCALAR IMPLICATURE, CLEFT, and PRESUPPOSITION judgments in adults, so these first two experiments develop that base-









Sentence Condition	Sentence	Congruency	Image	TVJT Exp. Resp	Rating Task Exp. Resp
Numeral baseline	‘There are two black ducks’	Congruent		–	3
		Incongruent		–	1
Scalar	‘Some of the ducks are black’	Congruent		True	3
		Incongruent		False	2
Presupposition	‘The duck is black’	Congruent		True	3
		Incongruent		False	2
It-cleft	‘It’s the duck that’s black’	Congruent		True	3
		Incongruent		False	2

Table 1: All conditions used in Experiments 1 and 2. The only difference between the experiments was the response options: ‘true’/‘false’ in Experiment 1 and 1/2/3 in Experiment 2. Experiment 1 did not include a numeral baseline condition.

line. The primary goal, however, is to compare two different tasks that have been used in similar acquisition studies: (i) a truth value judgment task (TVJT), and (ii) a 3-point rating task.

2.1 PARTICIPANTS. Participants were recruited via Amazon Mechanical Turk. To ensure high quality data, we only recruited individuals who had completed more than 1000 HITs with an acceptance rate of greater than 98%, currently resided in the U.S., self-reported being native speakers of English, and passed two attention check items within the HIT. Experiment 1 has data from 50 participants, and Experiment 2 has data from 48 participants.

2.2 MATERIALS. Experimental stimuli consisted of sentences using 15 different possible concrete nouns in Experiment 1, and 22 nouns (15 experimental + 7 baseline) in Experiment 2. The full list of 15 experimental nouns was chosen with the ultimate goal of presenting these stimuli to children as young as 4;0, so they were selected from the Peabody Picture Vocabulary Test Dunn & Dunn (2007). We chose only items that, according to this test, at least 90% of typically-developing four-year-olds would know, with the additional constraint that the words were concrete nouns that could also be easily depicted as solid-colored silhouette pictures. All images were taken from the free clip-art options on Microsoft Powerpoint.

Each noun was used in sentences that trigger (i) a scalar implicature, (ii) a presupposition of uniqueness, or (iii) an exhaustivity inference from an *it*-cleft. All sentences were presented with an image that is either CONGRUENT (i.e., ‘True’ or ‘3’) or INCONGRUENT (i.e., ‘False’ or ‘2’). Thus every noun had six possible conditions in which it could appear. The full experimental paradigm is shown for the noun *duck* in Table 1. Stimuli were separated into six lists, so that the same noun never appeared in the same list twice.

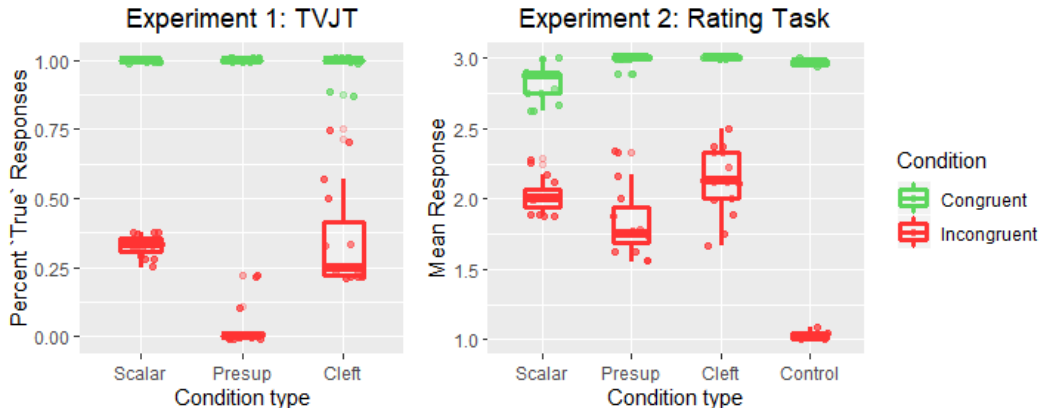


Figure 1: Results of experiments 1 and 2 showing the different pattern of responses depending on task type for the same stimuli. Each dot represents the mean response on a single item.

In the case of the SCALAR implicature condition, the INCONGRUENT sentence is infelicitous because *all* is more informative in this context. In the PRESUPPOSITION condition, the INCONGRUENT statement is infelicitous because there is more than one salient duck, so the presupposition of uniqueness from the definite determiner is violated. And in the CLEFT condition, the INCONGRUENT statement is infelicitous because there is another black thing in the image, making it non-exhaustive.

2.3 PROCEDURE. Participants began the experiment with instructions that told them to ‘judge whether what the sentence says is TRUE or FALSE about the picture’ (Experiment 1, TVJT) or to judge ‘how well the sentence goes with the picture, with 3 being good and 1 being not good’ (Experiment 2, Rating Task) In the TVJT, they saw one example of a true numeral baseline item; in the Rating Task, they see two examples: one of an incongruent numeral baseline item and one of a congruent numeral baseline item. All items were displayed on the same page. Participants had the option to change their initial answer, and there were no restrictions on the order in which participants answer the questions. In Experiment 1, we included two attention check items based on explicit numerals; any participant that missed one or both of these checks was excluded from analysis ($n = 3$ excluded, $n = 50$ included). In Experiment 2, we used the numeral baseline condition as inclusion criteria: any participant that did not get at least 6 out of the 7 items correct (where CONGRUENT = 3 and INCONGRUENT = 1) was excluded from analysis ($n = 7$ excluded, $n = 48$ included). The task took about two minutes to complete. Analysis was done in R (R Core Team 2019); we used the `tidyverse` suite Wickham (2017) for data manipulation, the `ez` package (Lawrence 2016) for analyses using an ANOVA, `ggplot2` (Wickham 2016) for data visualization.

2.4 RESULTS. Overall, we find that adults strongly distinguish between CONGRUENT and INCONGRUENT examples in all three experimental conditions (Figure 1). However, there are some striking differences in responses between the two experimental tasks.

In Experiment 1 (TVJT), participants chose ‘True’ at the expected rates (near or at 100% in all conditions). However, only in the PRESUPPOSITION condition did participants consistently select ‘False’ for the INCONGRUENT items. About a third of the time, participants respond with ‘True’ in the IT-CLEFT and SCALAR cases that were intended to be False. Despite

these differences between the conditions, independent samples t-tests shows that the differences between True and False items is highly significant in all conditions, all $ps < 0.01$.

In Experiment 2 (Rating task), independent samples t-tests shows that the differences between Congruent and Incongruent items is highly significant in all conditions, all $ps < 0.01$. There are also significant differences between conditions. We analyze these differences between conditions using ANOVAs with participants (F_1) and items (F_2) as random effects. Looking at the differences in INCONGRUENT items, the INCONGRUENT PRESUPPOSITION items are rated lower than INCONGRUENT IT-CLEFT [$F_1(1,47) = 17.88, p < 0.01$; $F_2(1,14) = 12.39, p < 0.01$] and SCALAR [$F_1(1,47) = 7.62, p < 0.01$; $F_2(1,14) = 3.92, p = 0.068$] items, but there is no significant difference between the INCONGRUENT IT-CLEFT and SCALAR items [$F_1(1,47) = 1.27, p = 0.265$; $F_2(1,14) = 2.46, p = 0.139$].

2.5 DISCUSSION. When only given two possible options in Experiment 1 (TVJT), are somewhat inconsistent in their responses, with acceptance rates of the infelicitous CLEFT and SCALAR items at 25% and 34%, respectively. These results are especially surprising given that there the stimuli contain only one item (one of the attention-check items) that is actually fully false. That is, the fact that the task is to determine True/False judgments, and that nearly all the remotely ‘bad’ examples are the ones that we expect adults to respond ‘False’ to, we have actually biased our data *towards* finding a stronger distinction between the True/False items than may actually be present. Though we might expect the difference to be exacerbated given this design, the pattern of responses that adults are showing reflects a much weaker distinction. By framing the question as a decision between ‘True’ and ‘False’, people may have responded in a more strictly *logical* way than they would have otherwise. In that sense, having determined that the INCONGRUENT items aren’t actually *false* (which they are not), just weird, then choosing ‘True’ is a reasonable choice for an adult.

The results from Experiment 2 (rating task) confirm that participants make the expected CONGRUENT/INCONGRUENT distinction, and that this distinction is different from the True/False distinction they made for the numeral baseline items. Participants consistently rated the INCONGRUENT items around a 2, showing that they make at least a 3-way distinction of congruent/incongruent/false. This distinction is not directly measurable in the design of Experiment 1, which gave only two options.

We have determined through these two experiments that, for our stimuli and primary question, the rating task used in Experiment 2 provides a more reliable measure of adult behavior compared to the TVJT used in Experiment 1. Though TVJT is a more common task in acquisition studies, we choose not to use it moving forward because the acceptability rating task is better suited to our design.

3. Experiment 3: Child rating task. We address the question of *it*-cleft exhaustivity development as a matter of *change over time* by looking at different age groups of children. If we see that either the SCALAR or PRESUPPOSITION responses are patterning with the CLEFT responses, then we take this as evidence for the development of a single mechanism that may be responsible for both inference types. We test children aged 4-6, a window of time when development of these inferences is both active and measurable (based on prior work).

3.1 PARTICIPANTS. Forty-three children completed this study: 15 4-year-olds (11 female), 20 5-year-olds (11 female), and 8 6-year-olds (2 female). An additional seven children began the study, but did not complete it. Eleven adult controls (6 female) completed the exact same task

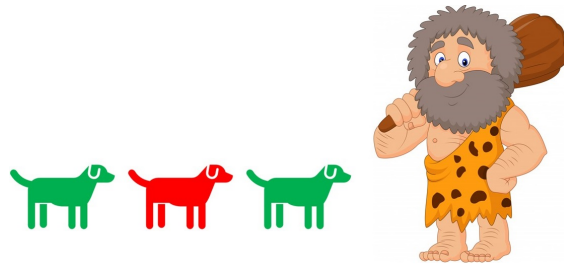


Figure 2: Example of a trial from Experiment 3 (child rating task). In this trial, Og says ‘Some of the dogs are green,’ an example of a congruent scalar implicature.

as the children.

3.2 MATERIALS. Stimuli and list creation were identical to the Experiment 2 rating task, but the sentence for each trial is presented as audio instead of text. Stimuli were recorded by a male speaker with a Mainstream American English accent. All practice trials had explicit numerals that make the item either True (and thus Congruent) or False.

3.3 PROCEDURE. The task design is based on the acceptability rating task used by Katsos & Bishop (2011). In our version of this task, children began the study by being introduced to Og³, the caveman. The experimenter explained that Og’s favorite food is strawberries, and that he is trying to learn to describe pictures. The child’s task was to reward Og based on how good of a job he does describing the pictures that he sees. Children were told that if Og does a good job, he should get three strawberries, but if he says something that’s silly, he should only get one. This explanation was followed by practice with either two or four training trials. If the child got the first two trials of training correct, they move on to the main experiment. If the child got one or both of the training trials incorrect, the experimenter provided another example of how to complete the task, again explaining the goal of rewarding Og when he does a good job. The child then saw two more such practice trials. All but two children passed the practice portion of the experiment, and those two that did not pass are not included in the analysis, though they were allowed to play the game with Og.

To help keep children’s interest, we also gave them three small plastic strawberries that they could use to ‘feed’ Og by hand. Most of the children chose to place the selected number of strawberries in a small bowl, though some of the children preferred to say a number aloud. For the main part of the experiment, children were seated in front of a computer screen on which Og and the pictures were displayed (see example of a screen in Figure 2) via PsychoPy (Peirce 2007). Each trial began with an image of Og next to a scene of three items he needed to describe. To get the child’s attention, the experimenter would direct their gaze to the screen, sometimes by asking a question about the items displayed and other times by saying something like ‘let’s see what Og says this time.’ When the child was ready, the experimenter played a short audio clip of the stimulus sentence. The child had the option to ask to hear it as many times as they wanted, though most only needed to hear the sentence one time. Once the child indicated the number of strawberries to reward Og with, the experimenter entered that number on the keyboard, and the child saw Og receive 1, 2, or 3 strawberries.

On several of the trials (at least one from each of the 8 conditions, including baselines),

³Image of Og was purchased from Vectorstock.com

Age	Condition	CONGRUENT	INCONGRUENT	Difference
4	NUM. CONTROL	2.77 (0.41)	1.17 (0.27)	1.60
	CLEFT	2.48 (0.62)	2.53 (0.65)	-0.05
	PRESUPPOSITION	2.76 (0.51)	2.18 (0.82)	0.58
	SCALAR	2.61 (0.61)	2.41 (0.63)	0.20
5	NUM. CONTROL	2.86 (0.29)	1.12 (0.15)	1.74
	CLEFT	2.84 (0.38)	2.55 (0.71)	0.30
	PRESUPPOSITION	2.79 (0.33)	2.15 (0.85)	0.64
	SCALAR	2.61 (0.67)	2.17 (0.80)	0.44
6	NUM. CONTROL	2.71 (0.49)	1.29 (0.32)	1.43
	CLEFT	2.60 (0.61)	1.93 (0.67)	0.67
	PRESUPPOSITION	2.81 (0.38)	1.69 (0.57)	1.12
	SCALAR	2.60 (0.56)	2.33 (0.75)	0.26
Adult	NUM. CONTROL	3.00 (0.00)	1.15 (0.22)	1.85
	CLEFT	2.82 (0.41)	1.83 (0.69)	0.99
	PRESUPPOSITION	2.95 (0.15)	1.58 (0.61)	1.38
	SCALAR	2.86 (0.23)	1.92 (0.70)	0.94

Table 2: Mean response scores for Experiment 3. Standard deviations are in parentheses. ‘Difference’ is computed as CONGRUENT minus INCONGRUENT.

the child was asked a follow-up question to get an idea of their reasoning. This prompt was phrased as ‘Can you tell me why you’re giving Og X strawberries?’ or just ‘Let’s tell Og why.’ Answers were recorded by hand by the experimenter/assistant or, if the child’s parents agreed, audio from the experiment was recorded throughout. Halfway through the experiment, Og took a break to dance, at which point the experimenter checked in with the child to see if they wanted to continue with more trials; all children agreed to continue playing the game.

3.4 RESULTS. Before looking at the results of the experimental conditions, we exclude several subjects based on their responses to the control items, as an inability to distinguish true vs. false sentences via the reward system of the game makes the responses on the experimental trials impossible to interpret. We operationalize a passing score on the control trials as being any participant who both rated the CONGRUENT (i.e., True) control items on average as greater than or equal to 2 and rated the INCONGRUENT (i.e., False) control items on average as less than 2. We exclude, before analysis, a total of 7 children (4 4-year-olds, 2 5-year-olds, and 1 6-year-old). All of the remaining participants distinguish the CONGRUENT and INCONGRUENT trials by at least 0.66 points, on average, though most distinguish them by well over 1.0 point. The mean ratings given in each age group for each condition are shown in Table 2.

Four- and five-year-old participants tend to over-accept the INCONGRUENT experimental items, even though they can accurately assign a low rating to clearly false items (numeral baseline). The 6-year-olds in this study have very adult-like interpretations of the PRESUPPOSITION and IT-CLEFT items, though overall they tended to over-accept the INCONGRUENT SCALAR items at rates comparable to the four- and five-year-olds. The aggregate results for

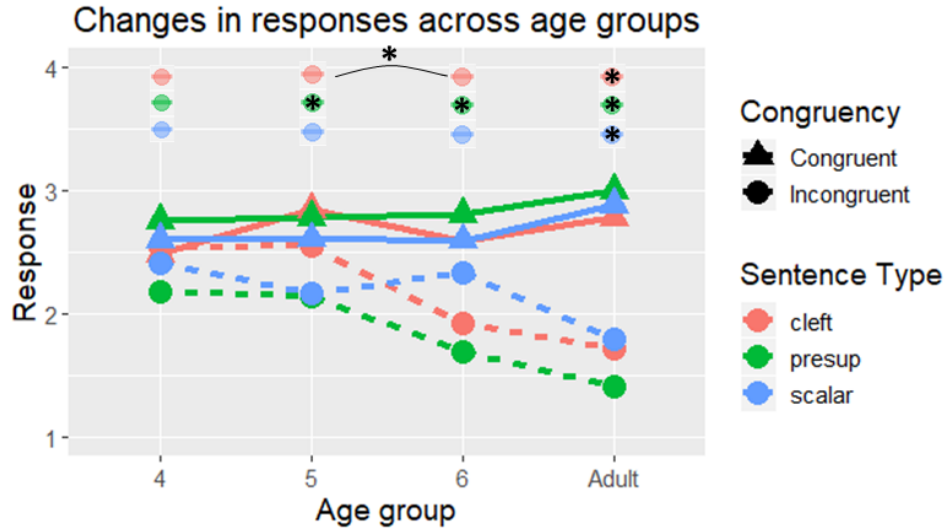


Figure 3: Responses to the CONGRUENT and INCONGRUENT items in each of the three experimental conditions across age groups. Within circles, * means significant or approaching significant difference between CONGRUENT & INCONGRUENT items. With liaison between the circles, * means a sig. difference on INCONGRUENT responses between age groups.

each age group are show in Figure 3. While PRESUPPOSITION judgments and, for the most part, IT-CLEFT judgments get smoothly more adult-like over time, the trajectory for SCALAR items is more delayed. Note that the number of subjects in the 6-year-old age group is rather small ($n=7$), though, so this trend may be due to noise.

Though the general trend is that children develop judgments that look more adult-like as they get older, this way of looking at the data does not fully address this study’s primary question: do children individually show *it*-clefts patterning together closely with either of the other two conditions? To get a sense of the individual trajectories, we conduct an exploratory follow-up analysis of the individual patterns of response by each participant to determine if there was a predominant pattern that emerged from what may be very heterogeneous data.

In order to break up the results into the predominant pattern, we identify a cut-off score for categorization. Within each of the three experimental conditions, if the participant has a mean difference of greater than 0.5 for the CONGRUENT - INCONGRUENT trials, we the participant as having an adult-like inference. This gives eight logical possibilities for patterns that could emerge. If the three relevant inferences develop fully independently (or rely on independent cognitive mechanisms) and are equally ‘difficult’, then we would expect that, for children with at least one of the inferences, response patterns will fall approximately equally across the seven remaining response patterns. The actual number of individuals whose responses fall into each of these patterns is shown in Table 3.

In order to assess the relationship between an individual participant’s responses on the different conditions, we analyze the difference score of CONGRUENT minus INCONGRUENT for each participant (where a score of 2 is the maximum difference between CONGRUENT and INCONGRUENT conditions, and 0 is no difference). This measure reflects the relative difference in acceptability that the participant assigns to sentences in each of the three experimental sentence conditions. We then compute the correlation between the three possible pairings of these

Category	Number of participants aged				Total
	4-yrs	5-yrs	6-yrs	Adult	
All Inferences	1	2	1	5	9
No Inferences	4	6	1	0	11
Only Cleft Inferences	1	1	0	0	2
Only Presupposition Inferences	2	1	1	1	5
Only Scalar Inferences	2	4	0	0	6
Presup & Cleft Inferences	0	2	2	2	6
Scalar & Cleft Inferences	0	0	1	1	2
Scalar & Presup Inferences	1	2	1	2	6
Total Ns	11	18	7	11	47

Table 3: Distribution of each of the eight possible patterns for participant responses, broken down by the number of participants in each age group who fall into each category. The highlighted rows show the two patterns that participants fall into at rates lower than is expected by chance.

Age	CLEFT-SCALAR	PRESUP-SCALAR	CLEFT-PRESUP
4	-0.067 ($p = 0.853$)	-0.116 ($p = 0.750$)	0.145 ($p = 0.690$)
5	-0.509 ($p = 0.031$)	0.139 ($p = 0.582$)	0.319 ($p = 0.198$)
6	0.036 ($p = 0.939$)	-0.639 ($p = 0.122$)	0.426 ($p = 0.341$)
Adult	0.485 ($p = 0.130$)	0.552 ($p = 0.078$)	0.469 ($p = 0.146$)

Table 4: Pearson’s correlation coefficient, r , and its p -value for each correlation.

conditions (CLEFT-SCALAR, PRESUPPOSITION-SCALAR, and CLEFT-PRESUPPOSITION) using these mean difference scores. The correlation coefficients and their corresponding significance levels are in Table 4, and these same values are visualized as best-fit lines in Figure 4.

None of the correlations in Figure 4 are significant above 0.05, though there are some clear general trends. In all cases, Adults show a medium-strength positive correlation between their responses on the two sentence conditions being compared; i.e., adults that more strongly reject one inference also more strongly reject the other two inferences. Notably, 4-year-olds show virtually no correlation in their responses on any of the comparisons, reflecting the fact that they are still developing all of the relevant inferences. The only comparison where participant responses have a consistent trend in becoming more strongly correlated across age groups is in the CLEFT-PRESUPPOSITION correlation. In the other two comparisons, correlations either remain weak or flip between a positive/negative trend for the child participants.

3.5 DISCUSSION. The overall pattern of responses is generally compatible with two possible interpretations: (i) presuppositions and *it*-cleft exhaustivity inferences share a common, indirect linguistic or other cognitive source, or (ii) *it*-cleft exhaustivity inferences are a completely separate phenomenon than implicature and presupposition judgments (as proposed by Križ (2015)). The patterns observed are inconsistent with an account of *it*-cleft exhaustivity that derives the inference from the same underlying mechanism as implicature judgments, as we observe these inferences patterning together less often than chance, and the two inferences

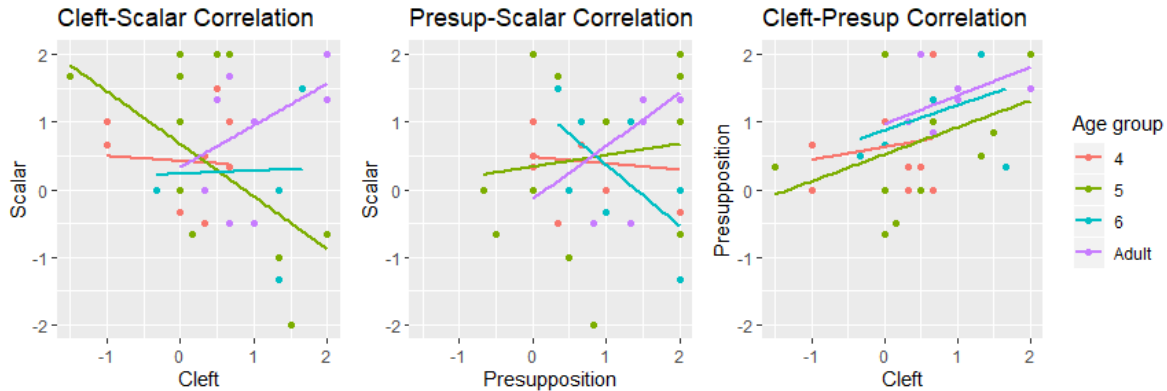


Figure 4: Correlations of mean difference scores between experimental conditions for each age group. Each dot represents a single participant.

are never even moderately positively correlated with each other during acquisition.

The aggregate results here function as a kind of sanity-check. We can clearly see that inferences become more adult-like over time, and this trend can be seen as essentially a replication of the main trends noted in several other acquisition studies. Four-year-olds, overall, over-accept many infelicitous utterances, while five year olds are beginning to make distinctions that looks adult-like. The crucial finding here, though, deals with the individual-level analysis.

4. General discussion. We investigated whether individual response patterns in children’s responses to quantity-related inferences supported an analysis where *it*-cleft exhaustivity is more similar to either semantic presupposition judgments or more pragmatic implicature calculations. We found that children’s responses to infelicitous uses of an *it*-cleft was positively correlated with their responses to presupposition violations, but not underinformative scalar implicatures. In development, children do not develop a more adult-like interpretation of English *it*-clefts alongside scalar items, making it unlikely that these two inferences share a common cognitive or linguistic mechanism. Children do, however, show a weak correlation between their development of adult-like judgments of non-exhaustive *it*-clefts and violations of a presupposition of uniqueness from singular definites, indicating that there may be a shared mechanism or reasoning used in both inferences.

Given that we do not find a evidence for a common source between *it*-cleft judgments and scalar implicature judgments, one may wonder if testing a different type of implicature would yield a different result. The choice to compare *it*-cleft exhaustivity to *scalar* implicatures, in particular, was driven by the existing proposals in the theoretical literature that tied these inferences together. However, the standard difference between scalar and non-scalar implicatures is that the scalar items have a lexical scale-mate. In the case of English *it*-clefts, though, there is no such scale-mate, and it’s not even clear which ‘scale’ the exhaustivity exists on. With these issues in mind, it’s reasonable to wonder whether exhaustivity inferences in clefts could still be related to implicatures, but rather to *non-scalar* implicatures. While we consider this possibility reasonable, we find it unlikely to have a considerable effect on the results, as the developmental trajectory for non-scalar implicatures may not be separate from that of the scalar cases. Katsos (2009) directly compared these two implicatures and found that on a quantitative measure, both children and adults responded to scalar and non-scalar kinds of implicatures in

consistent ways. though children's and adults' responses differed from each other.

A similar concern could be raised about the choice of presupposition trigger for this experiment. The full landscape of presupposition inferences and triggers is far too wide to make a sweeping conclusion about the relationship between 'presuppositions' and exhaustivity inferences from this study alone. In this study, we chose to keep the form of the stimuli maximally similar, while restricting the meaning to the more broad domain of 'quantity-related' inferences. However, one could also choose to keep meaning maximally similar and allow the structure to vary more, in which case a definite description such as 'the black thing is a duck' would have been a reasonable comparison. In keeping meaning maximally similar, it would be possible to test for the influence of extraneous factors such as information structure, as the order of the adjective and noun in 'the black thing is a duck' is the opposite of our *it*-cleft condition, 'it's the duck that's black,' though the meanings are identical.

Another potential issue that could arise from the choice of presupposition trigger in this experiment deals with covert domain restriction. Although we attempted to control the domain by presenting a simple scene of only three items, it is possible, for both the child and adult participants, that when they over-accepted the incongruent stimuli (particularly in the presupposition condition), they were able to "fix" the incongruity by adding an even smaller domain restriction within which the sentence would be congruent. While participants' use of this strategy could have added a certain amount of noise to the study, we note that the presupposition condition was likely the easiest for participants, in the sense that children developed towards adult-like judgments earlier compared to the other conditions and adults rated incongruent presupposition violations lower than they rated the other incongruent conditions.

Finally, we return to the possibility that a third type of inference, homogeneity, underlies the exhaustivity inference in *it*-clefts. The relationship between homogeneity and the French counterpart to *it*-clefts was studied by Tieu & Križ (2017) using a TVJT with children aged 3 to 6. Their crucial contrast used plural definites (e.g., 'The ducks are black', which gives the inference that the ducks are homogeneous with respect to being colored black – i.e., they're all black) and *it*-cleft structures in congruent and incongruent set-ups, very similar to our Experiment 3. While the results were potentially consistent with children's development of *it*-cleft exhaustivity relying on the same underlying inference as the plural definites, these two conditions did not clearly follow the same trajectory. However, the analysis used in this study looks at the data in the aggregate, so a direct comparison with the results of our individual analysis is not possible.

More recently in an extended version of this study, Tieu et al. (2019) use an individual-level analysis similar to our Table 3 to assess the possibility of a single cognitive/linguistic mechanism underlying both homogeneity inferences and implicature judgments. They find evidence supporting a distinction between the two inferences, at least in terms of their development, as they identified a group of children who had acquired homogeneity inferences while not yet having acquired adult-like judgments on scalar implicatures. Taking their findings together with our own, it is necessary to further investigate a homogeneity account of *it*-cleft exhaustivity, as this possibility is still consistent with the experimental findings discussed here. Though we leave this for future study, we note that a direct comparison of both singular and plural definites to *it*-cleft inferences is a clear next step in this line of research.

5. Conclusion. Through a within-subjects comparison of maximally-similar quantity-related inferences, we were able to assess the degree to which these inferences pattern together in their acquisition trajectory. We find evidence in support of an account where exhaustivity is not literally ‘built-on’ any other inferences, but rather may share a common source with presuppositions. The individual response patterns that we observe among 4-, 5-, and 6-year-olds offers a snapshot of the ways that this development can take place.

References

- Aravind, Athulya. 2018. *Presuppositions in context*: Massachusetts Institute of Technology dissertation.
- Aravind, Athulya, Martin Hackl & Ken Wexler. 2018. Syntactic and pragmatic factors in children’s comprehension of cleft constructions. *Language Acquisition* 25(3). 284–314.
- Barner, David, Neon Brooks & Alan Bale. 2011. Accessing the unsaid: The role of scalar alternatives in children’s pragmatic inference. *Cognition* 118(1). 84–93.
- Bill, Cory. 2015. Presuppositions vs. scalar implicatures in acquisition. *University of Pennsylvania Working Papers in Linguistics* 21(1). 3.
- Büring, Daniel & Manuel Križ. 2013. It’s that, and that’s it! exhaustivity and homogeneity presuppositions in clefts (and definites). *Semantics and Pragmatics* 6. 6–1.
- Chierchia, Gennaro, Stephen Crain, Maria Teresa Guasti, Andrea Gualmini, Luisa Meroni et al. 2001. The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures. In *Proceedings of the 25th boston university conference on language development*, 157–168.
- Cremers, Alexandre, Manuel Križ & Emmanuel Chemla. 2017. Probability judgments of gappy sentences. In *Linguistic and psycholinguistic approaches on implicatures and presuppositions*, 111–150. Springer.
- Cremers, Alexandre, Lyn Tieu & Emmanuel Chemla. 2016. Children’s exhaustive readings of questions. *Language Acquisition* .
- Destruel, Emilie. 2013. An empirical investigation of the meaning and use of the french c’est-cleft. (*Unpublished doctoral dissertation*). *University of Texas at Austin, Austin, TX* .
- Destruel, Emilie & Bryan Donaldson. 2017. Second language acquisition of pragmatic inferences: Evidence from the french c’est-cleft. *Applied Psycholinguistics* 38(3). 703–732.
- Destruel, Emilie, Daniel Velleman, Edgar Onea, Dylan Bumford, Jingyang Xue & David Beaver. 2015. A cross-linguistic study of the non-at-issueness of exhaustive inferences. In *Experimental perspectives on presuppositions*, 135–156. Springer.
- DeVaugh-Geiss, Joseph P, Swantje Tönnis, Edgar Onea & Malte Zimmerman. 2017. An experimental investigation of (non-)exhaustivity in clefts. In *Proceedings of Sinn und Bedeutung 21*, .
- DeVaugh-Geiss, Joseph P, Malte Zimmermann, Edgar Onea & Anna-Christina Boell. 2015. Contradicting (not-) at-issueness in exclusives and clefts: An empirical study. In *Semantics and linguistic theory*, vol. 25, 373–393.
- Diessel, Holger & Michael Tomasello. 2001. The acquisition of finite complement clauses in english: A corpus-based analysis. *Cognitive Linguistics* 12(2). 97–141.
- Dunn, Lloyd M. & M. Dunn, Douglas. 2007. *The peabody picture vocabulary test fourth edition* 4th edn.

- Foppolo, Francesca, Maria Teresa Guasti & Gennaro Chierchia. 2012. Scalar implicatures in child language: Give children a chance. *Language Learning and Development* 8(4). 365–394.
- Grice, H Paul. 1975. Logic and conversation. In P. Cole & J.L. Morgan (eds.), *Syntax and semantics, vol. 3, speech acts*, 41–58. Academic Press, New York.
- Katsos, Napoleon. 2009. Evaluating under-informative utterances with context-dependent and context-independent scales: experimental and theoretical implications. *Experimental Semantics and Pragmatics* 51–73.
- Katsos, Napoleon & Dorothy VM Bishop. 2011. Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition* 120(1). 67–81.
- Križ, Manuel. 2015. Homogeneity, non-maximality, and all. *Journal of Semantics* 33(3). 493–539.
- Križ, Manuel. 2016. Referentiality, exhaustivity, and trivalence in it-clefts. *Ms. LSCP*.
- Lawrence, Michael A. 2016. *ez: Easy analysis and visualization of factorial experiments*. <https://CRAN.R-project.org/package=ez>. R package version 4.4-0.
- Noveck, Ira A. 2001. When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition* 78(2). 165–188.
- Papafragou, Anna & Julien Musolino. 2003. Scalar implicatures: Experiments at the semantics–pragmatics interface. *Cognition* 86(3). 253–282.
- Peirce, JW. 2007. PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods* 162. 8–13.
- R Core Team. 2019. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. <http://www.R-project.org/>.
- Skordos, Dimitrios & Anna Papafragou. 2016. Children’s derivation of scalar implicatures: Alternatives and relevance. *Cognition* 153. 6–18.
- Smith, Carol L. 1980. Quantifiers and question answering in young children. *Journal of Experimental Child Psychology* 30(2). 191–205.
- Tieu, Lyn & Manuel Križ. 2017. Connecting the exhaustivity of clefts and the homogeneity of plural definite descriptions in acquisition. In *BUCLD 41 proceedings*, .
- Tieu, Lyn, Manuel Križ & Emmanuel Chemla. 2019. Children’s acquisition of homogeneity in plural definite descriptions. *Frontiers in psychology* 10. 2329.
- Velleman, Dan, David Beaver, Emilie Destruel, Dylan Bumford, Edgar Onea & Liz Coppock. 2012. It-clefts are IT (inquiry terminating) constructions. In *Semantics and linguistic theory*, vol. 22, 441–460.
- de Villiers, Jill & Tom Roeper. 2016. The acquisition of complements. In *The oxford handbook of developmental linguistics*, .
- Wickham, Hadley. 2016. *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>.
- Wickham, Hadley. 2017. *tidyverse: Easily install and load the 'tidyverse'*. <https://CRAN.R-project.org/package=tidyverse>. R package version 1.2.1.