Scalar quantifiers: Logic, acquisition, and processing

Bart Geurts
Napoleon Katsos
Chris Cummins
Jonas Moons
Leo Noordman

Abstract

Superlative quantifiers ("at least 3", "at most 3") and comparative quantifiers ("more than 2", "fewer than 4") are traditionally taken to be interdefinable: the received view is that "at least n" and "at most n" are equivalent to "more than n–1" and "fewer than n+1", respectively. Notwithstanding the prima facie plausibility of this claim, Geurts and Nouwen (2007) argue that superlative quantifiers have essentially richer meanings than comparative ones. Geurts and Nouwen's theory makes three kinds of predictions that can be tested by experimental means. First, it predicts that superlative and comparative quantifiers should give rise to different patterns of reasoning. Secondly, it leads us to expect that children will master comparative quantifiers before superlative ones. Thirdly, superlative quantifiers should be harder to process than comparative ones. We present three experiments that confirm these predictions.

January 30, 2009

1. Introduction

The study of quantifying statements like "All A are B" and "Some A are B" has a long history in logic, linguistics, and psychology, and despite all differences between and within these disciplines, there is a *de facto* standard doctrine as to how quantifying expressions are to be interpreted. According to this doctrine, the meaning of a quantifying statement of the form "Q A are B", where Q is a quantifier, can always be given in terms of a relation between the set of A's and the set of B's. For example, "All lawyers are crooks" means that the set of lawyers is a subset of the set of crooks, and "Some lawyers are crooks" means that there is a non-empty set of lawyers who are crooks; Table 1 gives more examples of this style of analysing quantification, which is prevalent in logic and linguistics (Barwise and Cooper 1981), the psychology of language (Moxey and Sanford 1993), and the psychology of reasoning (Evans et al. 1993).[1]

sentence	*interpretation*
All A are B	$[\![A]\!] \subseteq [\![B]\!]$
Some A are B	$[\![A]\!] \cap [\![B]\!] \neq \varnothing$
No A are B	$[\![A]\!] \cap [\![B]\!] = \varnothing$
More than 2 A are B	$\text{card}([\![A]\!] \cap [\![B]\!]) > 2$
At least 3 A are B	$\text{card}([\![A]\!] \cap [\![B]\!]) \geq 3$
Fewer than 3 A are B	$\text{card}([\![A]\!] \cap [\![B]\!]) < 3$
At most 2 A are B	$\text{card}([\![A]\!] \cap [\![B]\!]) \leq 2$

Table 1: The standard interpretation of quantifier expressions. Notation: $[\![X]\!]$ is the set of all X's (in a given context); card(X) is the cardinality of set X.

Recently, researchers in semantics and pragmatics have begun to argue that the standard model isn't rich enough for handling all quantifiers. One case in point is Ariel's (2004) work on "most"; another is Geurts and Nouwen's (2007) theory of scalar quantifiers, which is our starting point in this paper. In the following we will first outline Geurts and Nouwen's theory, and then present three experiments based on it. Contrary to the standard model, G&N's analysis predicts that the logic of quantifiers like "at least/most 3" is different from that of quantifiers like "more/fewer than 3", which is to say that arguments which are valid when formulated

1. A prominent exception to the standard line is Chater and Oaksford's (1999) probabilistic interpretation of quantification; see Geurts (2003) for critical discussion of this approach. The classical logical treatment of "some" and "all" is not an exception, because it is equivalent to what we call the standard account.

in terms of the latter need not remain valid when recast in terms of the former. This prediction was tested in an offline study, which is Experiment 1 of this paper. A further prediction of the G&N theory is that quantifiers of the "at least/most" type are more complex than quantifiers of the "more/fewer" type, which would lead us to expect that the acquisition of the former will lag behind the latter, and that "at least" and "at most" are harder to process than "more than" and "fewer than". These predictions were tested in Experiments 2 and 3.

2. Scalar quantifiers

Scalar quantifiers come in two types: superlative ("at least 3", "at most 3") and comparative ("more than 2", "fewer than 4"). As shown in Table 1, these two types are standardly taken to be interdefinable: since "$m \geq n$" and "$m \leq n$" are equivalent to "$m > n$–1" and "$m < n$+1", respectively, it follows from the definitions in Table 1 that "at least n" and "at most n" are equivalent to "more than n–1" and "fewer than n+1", respectively. Prima facie, this seems entirely plausible, but G&N argue that these equivalences don't hold. For example, G&N observe that, if "at most 3 martinis" was equivalent to "fewer than 4 martinis", the following should be equally acceptable, which they are not:

?Berta didn't have at most 3 martinis.
Berta didn't have fewer than 4 martinis.

In order to account for these and other observations, G&N argue that, while comparative quantifiers can retain their conventional meaning (as per Table 1), the meanings of superlative quantifiers are richer than the standard model would have it. More concretely, G&N propose the following interpretations for "at least" and "at most":

- "At least n A are B" means that the speaker
  – is *certain* that there is a set of n A's that are B, and
  – considers it *possible* that there is a larger set of A's that are B.
- "At most n A are B" means that the speaker
  – considers it *possible* that there is a set of n A's that are B, and
  – is *certain* that there is no larger set of A's that are B.

On G&N's account, superlative quantifiers introduce a modal element which is absent from the meanings of comparative quantifiers. It is this element of modality which explains, according to G&N, why "Berta didn't have at most 3 martinis" is infelicitous: it is because there are restrictions

on the occurrence of (epistemic) modals in the scope of negation, as is illustrated by the fact that "Berta didn't have maybe 3 martinis" is just as bad, while "Berta had maybe 3 martinis" is fine.

G&N's theory makes three kinds of predictions that can be tested by experimental means. To begin with, it predicts that (i) the two types of quantifiers should give rise to different inference patterns. Furthermore, if it is true that superlative quantifiers ("at least/most") are more complex than comparative ones, they should (ii) take longer to learn and (iii) be harder to process. In the remainder of this paper, we discuss these predictions in some detail, and present three experimental studies designed to test whether the differences predicted by the theory are borne out by the facts.

3. Inference patterns

According to G&N's theory, a valid argument need not remain valid if a comparative quantifier is replaced with its superlative counterpart, and we should expect these differences to be reflected in people's reasoning. To illustrate, while on the standard account the following arguments are both valid, only the first one is valid on G&N's analysis:

Berta had 3 beers ⇒ Berta had more than 2 beers.
Berta had 3 beers ⇒ Berta had at least 3 beers.

The meaning of "Berta had at least 3 beers", according to G&N, is that the speaker (i) is certain that Berta had 3 beers and (ii) considers it possible that she had more. So a speaker who claims that Berta had 3 beers and at least 3 beers flatly contradicts himself: if Berta had 3 beers, it is not (epistemically) possible that she had more.

Another contrast between the standard theory and G&N's is illustrated by the following pair of arguments:

Berta had fewer than 3 beers ⇒ Berta had fewer than 4 beers.
Berta had at most 2 beers ⇒ Berta had at most 3 beers.

As before, while the standard model predicts that both arguments are valid, G&N's model predicts this only for the first argument. If someone says that Berta had "at most 2 beers", then according to G&N he rules out the possibility that she had 3 beers, whereas saying that Berta had "at most 3 beers" implies that this *is* a possibility; so on G&N's analysis the second argument is not valid.

Things get slightly more complicated with arguments like the following:

Berta had 3 beers ⇒ Berta had fewer than 4 beers.
Berta had 3 beers ⇒ Berta had at most 3 beers.

According the standard model, both arguments are valid, and since G&N accept the standard analysis of "fewer than", their theory, too, predicts that the first one is valid, which seems reasonable enough. What about the second argument? On G&N's analysis, the meaning of "Berta had at most 3 beers" is that that the speaker (i) considers it possible that Berta had 3 beers, but (ii) rules out the possibility that she had more. Logically speaking, this follows from the premiss that Berta had 3 beers, but on a pragmatic level there is a tension between the premiss and the first component of the conclusion: arguably, if it is given that Berta had 3 beers, it is not merely possible but certain that she had 3 beers.

Suppose someone who believes that sentence S *is* or *must be* true is asked whether S *might* be true. How would he respond? Experimental evidence suggests that the outcome will be mixed. In the context of an acquisition study, Noveck (2001) asked adult controls whether there *might* be a parrot in a box that was known to contain a parrot. In one experiment, 35% of participants said there might be, while in another the percentage was as high as 75%, but still 25% short of unanimity.[2] Similarly, in an unpublished study, we asked participants to imagine Berta saying: "I'm certain that Vernon is drunk", after which they had to decide whether this implies that Berta believes that Vernon *might* be drunk. 61% of the participants said it did.

Although there is a great deal of variation between these studies, they agree that people's responses to this type of task are mixed. Assuming this much is right, G&N's analysis of "at most" leads us to expect that people's responses to the second of the arguments above will likewise be mixed.

4. Experiment 1: Logic

In our first experiment, we tested the predictions discussed in the last section, using a paper-and-pencil design.

Participants

28 students of management science at the University of Nijmegen, all native speakers of Dutch, were paid for participating in this experiment.

2. The main difference between the two experiments was that the latter included a more elaborate training session.

Their ages ranged between 18 and 24 years; their mean age was 19;8 years. 16 of the participants were male.

Materials and procedure

Each participant received a 48-page questionnaire, in Dutch, with instructions printed on the front cover.[3] Each page in the questionnaire presented a single-premiss argument, and participants had to indicate whether the conclusion followed from the premiss by ticking a box. Sample materials are given in Table 2. The content words as well as the proper names in the materials varied between conditions: there were three beverages (beer, wine, and lemonade) and six proper names (three male, three female), which were randomly assigned to conditions. Every participant saw the arguments in a different, randomised order.

	premiss	*conclusion*
(1) a.	Berta had 3 beers	Berta had at least 3 beers
b.	Berta had 3 beers	Berta had more than 2 beers
(2) a.	Berta had 3 beers	Berta had at most 3 beers
b.	Berta had 3 beers	Berta had fewer than 4 beers
(3) a.	Berta had at most 2 beers	Berta had at most 3 beers
b.	Berta had fewer than 3 beers	Berta had fewer than 4 beers
(4) a.	Berta had at least 3 beers	Berta had 3 beers
b.	Berta had at most 3 beers	Berta had 3 beers
(5) a.	Berta had 3 or 4 beers	Berta had at least 3 beers
b.	Berta had 2 or 3 beers	Berta had at most 3 beers

Table 2: Materials used in Experiment 1. (1a), (2a), and (3a) were the critical items; the remainder were controls.

The critical items in this experiment were (1a), (2a), and (3a). As explained in the last section, these are the arguments on which G&N's logic deviates from that of the standard theory: while the latter predicts that these inferences should all go through, the former predicts that they shouldn't. Items (1b), (2b), and (3b) served as direct controls for the critical arguments. The remaining items served a threefold purpose: to draw attention away from the critical items, to gauge the complexity of this type of reasoning task, and to explore possible interactions between scalar quantifiers and various linguistic environments, like disjunction, for example, as illustrated by items (5a,b).

3. The Dutch expressions used for the quantifiers were "minstens" ("at least"), "hoogstens" ("at most"), "meer dan" ("more than"), and "minder dan" ("fewer than").

	premiss	*conclusion*	% (*sd*)
(1) a.	Berta had 3 beers	Berta had at least 3 beers	50 (51)
b.	Berta had 3 beers	Berta had more than 2 beers	100
(2) a.	Berta had 3 beers	Berta had at most 3 beers	61 (50)
b.	Berta had 3 beers	Berta had fewer than 4 beers	93 (26)
(3) a.	Berta had at most 2 beers	Berta had at most 3 beers	14 (36)
b.	Berta had fewer than 3 beers	Berta had fewer than 4 beers	71 (46)
(4) a.	Berta had at least 3 beers	Berta had 3 beers	50 (51)
b.	Berta had at most 3 beers	Berta had 3 beers	18 (39)
(5) a.	Berta had 3 or 4 beers	Berta had at least 3 beers	96 (19)
b.	Berta had 2 or 3 beers	Berta had at most 3 beers	93 (26)

Table 3: Percentages of participants who accepted as valid the arguments in Experiment 1.

Results and discussion

The main results of this experiment are presented in Table 3. Setting aside the critical items (on which the standard theory and G&N's disagree) and two items which gave rise to unexpected interpretations (see below), responses were correct in 87% of the cases, which goes to show that this type of task is not particularly difficult. In the following we will report only on the control items that are relevant for adjudicating between the two theories.

Our main finding was that there were significant differences between (1a), (2a), and (3a), on the one hand, and (1b), (2b), and (3b), on the other ($p < .0001$, .005, and .0001, using McNemar's test): in these cases, arguments with comparative quantifiers were endorsed more often than their superlative counterparts. While the standard theory doesn't predict these contrasts, G&N's account does. However, there is one wrinkle in the data: although the response rates for (1a) and (1b) were significantly different, as predicted, it was unexpected that the argument in (1a) should still be endorsed at a fairly high rate. We believe this is because, in the context of arguments involving the quantifier "at least 3 beers", some participants adopted an "at least" interpretation of the expression "3 beers"; so in these cases "3 beers" is interpreted as "3 or more beers". This explanation is confirmed by the finding that when the argument in (1a) was reversed, 50% of the participants said that the conclusion was valid; in the corresponding argument with "at most" we didn't see such an effect; cf. item (4b).[4] Additional support for this explanation was obtained in

4. See Geurts (2006) for an analysis of number words which predicts that, while the dom-

a follow-up study, in which we compared plain "3 beers" with "exactly 3 beers", expecting that the latter would reduce the rates of positive responses. This turned out to be the case, as is shown in Table 4.

premiss	*conclusion*	% (*sd*)
Berta had 3 beers	Berta had at least 3 beers	58 (50)
Berta had exactly 3 beers	Berta had at least 3 beers	21 (42)
Berta had at least 3 beers	Berta had 3 beers	58 (50)
Berta had at least 3 beers	Berta had exactly 3 beers	0

Table 4: Results of a follow-up study ($n = 24$). The differences between the arguments with "3 beers" and their counterparts with "exactly 3 beers" were both significant ($p < .02$ and $p < .0001$, respectively).

Note, finally, the contrast between the response rates for the arguments in (1a) and (5a). On the face of it, the only difference between these arguments is that the premiss of first argument is stronger (i.e. more informative). Nevertheless, whereas the second argument was accepted by practically all participants, the first argument was rejected half of the time. This finding makes little sense from the standard point of view, but is readily explained by G&N's theory. The sentence "Berta had 3 or 4 beers" conveys that the speaker is certain that Berta had 3 beers and considers it possible that she had more than 3 beers, and according to G&N's analysis this is precisely what the conclusion says. [5]

To conclude, the results of this experiment indicate that the way people reason with scalar quantifiers is more in line with G&N's theory than with the standard view.

5. Complexity

The most straightforward prediction made by G&N's theory is that the meanings of superlative quantifiers are more complex than those of their comparative counterparts. To see how this follows, consider how G&N analyse the meanings of "fewer than" and "at most":

inant interpretation of "3 beers" is exact, there is a recessive "at least" construal, as well.

5. A cursory inspection of the percentages in Table 3 suggests that the results for the quantifiers "at most" and "fewer than" are less crisp than for "at least" and "more than". We believe that this impression is correct, and that the difference is due to the fact that "downward entailing" quantifiers (like "at most" and "fewer than") are harder to process than "upward entailing" quantifiers (like "at least" and "more than"). We will return to this topic in the next section.

- "Fewer than n A are B" means that according to the speaker the number of B's is larger than the number of A's.
- "At most n A are B" means that the speaker
  – considers it *possible* that there is a set of n A's that are B, and
  – is *certain* that there is no larger set of A's that are B.

According to these definitions, both quantifiers involve comparing set sizes, but whereas the meaning of "Fewer than n A are B" merely says that one set is larger than the other, "At most n A are B" entails two claims about the A's and the B's, and furthermore these claims essentially involve different degrees of certainty. If this semantic analysis is on the right track, it is practically inevitable that the cognitive representation of "at most" should be more complex than that of "fewer than". The same, mutatis mutandis, "at least" and "more than".

If superlative quantifiers are inherently more complex than comparative ones, we should expect that, at some point in the course of language learning, children will have problems with superlative but not with comparative quantifiers. Experimental data reported by Musolino (2004) confirm that this is the case. Musolino presented 5-year-old children with collections of cards showing varying numbers of stars or smiley faces. In each trial, children had to select the cards meeting a description like "cards with more than 2 stars", "cards with at most 2 smiley faces", and so on. As shown in Table 5, children had no problems with "exactly 2" and "more than 2", while with superlative quantifiers they performed at chance level, thus corroborating G&N's theory. (Unfortunately, Musolino's materials didn't include items with "fewer than".)

expression	%
cards with exactly 2 {stars/smiley faces}	100
cards with more than 2 {stars/smiley faces}	88
cards with at least 2 {stars/smiley faces}	54
cards with at most 2 {stars/smiley faces}	50

Table 5: Percentages of correct responses in Musolino's (2004) experiment.

The aim of our second experiment was to follow up on Musolino's findings, while in our third experiment we wanted to see whether adults, too, find superlative quantifiers harder to process than comparative ones. In the design of these experiments, we had to take into account that, in addition to the superlative/comparative distinction, there is at least one further factor that may affect the complexity of a quantifier: quantifiers

may be either upward or downward entailing, and there is evidence that upward entailing quantifiers are easier to process than downward entailing ones. Upward entailing quantifiers license inferences from sets to supersets. For example, if there are some *mice* in the cupboard, then it must be true that there are some *rodents* in the cupboard. Similarly:

There are more than 2 *spades* in the deck
⇒ There are more than 2 *cards* in the deck.
There are at least 3 *spades* in the deck
⇒ There are at least 3 *cards* in the deck.

So, "more than" and "at least" are upward entailing. By contrast, "fewer than" and "at most" are downward entailing, since they license inferences in the opposite direction, i.e. from sets to subsets:

There are fewer than 3 *cards* in the deck
⇒ There are fewer than 3 *spades* in the deck.
There are at most 2 *cards* in the deck
⇒ There are at most 2 *spades* in the deck.

It has been known for some time that upward and downward entailment are important properties from a semantic point of view (e.g., Ladusaw 1979, Barwise and Cooper 1981, van der Wouden 1997). Also, they are properties that children master at a very early stage (Gualmini 2004), and, most importantly for our current purposes, they affect the complexity of linguistic expressions. Intuitively speaking, upward and downward entailing expressions are positive and negative, respectively (negative expressions like "not" are downward entailing), and therefore it doesn't come as a surprise that downward entailing expressions are more difficult to process than upward entailing ones. For example, Just and Carpenter (1971) found, in a sentence verification task, that statements with upward entailing quantifiers have shorter response latencies, and Geurts and van der Slik (2005) report on a reasoning experiment in which arguments with "at least" or "more than" gave rise to fewer errors than their counterparts with "at most" or "fewer than" (see also Geurts 2003).

Putting together these observations with G&N's analysis of scalar quantifiers (cf. Table 6), we should expect that downward entailing quantifiers are harder to process and acquired later than upward entailing ones, and that superlative quantifiers are harder to process and acquired later than comparative ones. Therefore, "more than", which is comparative and upward entailing, should be easier to process than, and acquired before, any of the others, while "at most", which is superlative and downward entailing, should be the hardest to process, and acquired last.

	comparative	*superlative*
upward entailing	more than	at least
downward entailing	fewer than	at most

Table 6: Two factors predicted to affect the complexity of a quantifier: entailment and scalar type.

It may be observed that, while these predictions do not follow from the standard view on scalar quantifiers, they are compatible with it. According to the standard doctrine, "at least *n*" and "at most *n*" are equivalent to "more than *n*–1" and "fewer than *n*+1", respectively. That is to say, "at least *n*" and "at most *n*" *express the same information as* "more than *n*–1" and "fewer than *n*+1", respectively. If we wanted to stick to the standard theory, we might hypothesise that, although in terms of information content comparative and superlative quantifiers cannot be differentiated, their cognitive representations are different. For instance, it might be supposed that, in the mental lexicon, "at least *n*" and "at most *n*" are defined as "more than *n*–1" and "fewer than *n*+1", respectively. Obviously, with this auxiliary hypothesis in place, the standard theory predicts the same complexity profile for scalar quantifiers as does G&N's.

However, even if the predictions made by the standard theory can be made to partially converge with G&N's, it should be noted, first, that the auxiliary hypothesis that makes this possible is ad hoc: the standard view is equally consistent with the assumption that "more than" and "fewer than" are defined in terms of "at least" and "at most", and it doesn't imply in any way that it should be the other way round. Secondly, even with this auxiliary hypothesis, the standard theory fails to explain the results of Experiment 1 as well as the linguistic evidence presented by Geurts and Nouwen (2007).

The upshot of these considerations is that the following experiments are better seen as a test of G&N's theory than of the standard view on scalar quantifiers, since whatever the outcome of these experiments will be, it will be consistent with the standard view.

6. Experiment 2: Acquisition

As discussed in the foregoing, Musolino's (2004) acquisition study provides evidence for our claim that superlative quantifiers are mastered later than comparative ones. However, since Musolino used an incomplete paradigm of scalar quantifiers (he didn't have items with "fewer

than"), and his participants performed at chance level with both superlative quantifiers, his results show less differentiation than our theory predicts. In order to obtain finer-grained results, we completed the paradigm and turned to older children: whereas Musolino's participants were 5-year-olds, ours were 11-year-olds.

We used an action-based task modeled after Pouscoulous et al.'s (2007) Experiment 3. Participants were presented with an array of six boxes and six small toys. The boxes and the toys were laid out in front of each participant in one of three arrangements, where two, three or four of the boxes already contained a toy (we will refer to these as "2-, 3-, and 4-arrangements"); the remaining toys were put on the table. Participants were told that the experimenter would utter a sentence and they would have "to make the boxes and toys match the sentence". They were also told that they could add toys to the boxes, remove toys from the boxes, or leave everything as it was.

This study was part of a larger experiment using the same methodology, in which participants were presented with three blocks of quantifier sentences in total. The first block included sentences with numerals (e.g., "Two/Four/Six of the boxes have a toy"), the second block contained quantified sentences ("All/None/Some of the boxes have a toy"), and the third block contained sentences with "exactly three" and the scalar quantifiers; the first two blocks are left out of account here. However, we should mention that all participants performed flawlessly with the numerals in the first block, which indicates that they had a good understanding of number words and had no problems with the experimental task as such.

Participants

There were two participant groups of native speakers of British English: 35 normally-developing 10- and 11-year-old children were recruited from a primary school (18 female, mean age 10;8, range 10;2-11;5) and 35 adults were recruited from the student pool of the University of Cambridge (22 female, mean age 22;3, range 19;1-24;3).

Materials and procedure

The materials used in this experiment were the following:

Exactly 3 boxes have a [toy]
At least 3 boxes have a [toy]
At most 3 boxes have a [toy]

More than 3 boxes have a [toy]
Fewer than 3 boxes have a [toy]

Each participant saw these sentences in one of two possible orders, with the "exactly 3" sentence being the first in both cases. The order of presentation of the 2-, 3-, and 4-arrangements was randomised.

On each trial, participants were presented with one sentence, and when they had performed an action to make the sentence true, the experimenter would restore the original arrangement and proceed to the next sentence, without giving feedback. Once participants had heard all the sentences in one arrangement, the next arrangement was laid out, and the same sentences were presented in the same order. While the boxes remained the same throughout the experiment, new toys were brought in for each arrangement in order to keep the younger participants interested.

Results and discussion

While our adult participants performed flawlessly across the board, children's responses were more variable. For each sentence/arrangement pair, Table 7 lists, for each of seven possible actions, the percentage of 11-year-olds that chose that action; the correct actions are shaded. Furthermore, for each sentence type, the rightmost column of Table 7 gives the percentage of children that performed correctly in all three arrangements.

It may be recalled that children as well as adults performed perfectly on the numeral items of an experiment preceding the one presented here. Likewise, in the current experiment, the children were perfect with "exactly 3", which is in line with Musolino's results. Furthermore, the 11-year-olds in our experiment also performed at ceiling level with "more than", where Musolino's 5-year-olds responded correctly 88% of the time. It can hardly be doubted, therefore, that the children in our study had a good understanding of number words and no problems with the task requirements as such.

Unlike Musolino, we obtained different rates of correct responses for "at least" and "at most". Musolino's 5-year-olds were at chance level with both superlative quantifiers; the 11-year-olds in our experiment did very well with "at least" (88% correct), but apparently "at most"was too difficult even for this age group (43% correct). Finally, with "fewer than" sentences, which were lacking from Musolino's materials, children responded correctly 77% of the time. Hence, it would seem that these data are in line with our predictions: "at most" is the hardest of the scalar

	2-arrangement		*3-arrangement*		*4-arrangement*		
	response	%	*response*	%	*response*	%	*% all correct (sd)*
exactly 3	0	0	0	100	0	0	100
	+1	100	+1	0	+1	0	
	+2	0	+2	0	+ 2	0	
	+3	0	+3	0	–1	100	
	+4	0	–1	0	–2	0	
	–1	0	–2	0	–3	0	
	–2	0	–3	0	–4	0	
% correct		100		100		100	
at least 3	0	11	0	71	0	46	88 (35)
	+1	57	+1	20	+1	6	
	+2	29	+2	3	+2	3	
	+3	3	+3	0	–1	37	
	+4	0	–1	6	–2	9	
	–1	0	–2	0	–3	0	
	–2	0	–3	0	–4	0	
% correct		89		94		91	
at most 3	0	23	0	46	0	29	43 (50)
	+1	40	+1	29	+1	23	
	+2	37	+2	11	+2	3	
	+3	0	+3	0	–1	37	
	+4	0	–1	14	–2	9	
	–1	0	–2	0	–3	0	
	–2	0	–3	0	–4	0	
% correct		63		60		46	
more than 3	0	0	0	0	0	80	97 (17)
	+1	3	+1	80	+1	17	
	+2	83	+2	17	+2	3	
	+3	11	+3	3	–1	0	
	+4	3	–1	0	–2	0	
	–1	0	–2	0	–3	0	
	–2	0	–3	0	–4	0	
% correct		97		100		100	
fewer than 3	0	80	0	6	0	3	77 (42)
	+1	0	+1	6	+1	0	
	+2	9	+2	6	+2	3	
	+3	0	+3	0	–1	14	
	+4	0	–1	83	–2	80	
	–1	11	–2	0	–3	0	
	–2	0	–3	0	–4	0	
% correct		91		83		80	

Table 7: Percentages of times 11-year-olds chose each of the possible actions in Experiment 3. Legend: "+n" = add n toys; "–n" = remove n toys; "0" = leave as is. Correct responses are shaded. The right-most columns gives, for each quantifier, the percentage of participants who performed correctly in all three arrangements.

quantifiers, "more than" is the easiest, and "at least" and "fewer than" fall between these two.

In order to test this impression, the data from the 11-year-old children were analysed using Cochran's Q-test, which is suitable for non-parametric frequency data of a dichotomous nature. The analysis revealed that there were significant differences between the four means (Q = 37.89; $p < .001$). Further comparisons using McNemar's test showed that children's performance with the upward-entailing modifiers, "more than" and "at least", was better than with their downward-entailing counterparts, "fewer than" and "at most" ($p < .012$ and $p < .001$, respectively). The differences between comparative and superlative quantifiers reached statistical significance for the comparison between "fewer than" and "at most" ($p < .01$) though not for the comparison between "more than" and "at least" ($p = .12$).

In order to compare 11-year-olds' responses with those of the adult controls, we applied McNemar's test to each of the quantifier conditions. This yielded significant differences for "at most" and "fewer than" ($p < .005$ in both cases) and a marginally significant difference in the case of "at least" ($p = .063$). As expected, children did not differ from adults in the "more than" condition ($p = 1$, ns).

The key predictions we derived from Geurts and Nouwen's theory of scalar quantifiers and extant experimental data on upward and downward entailment were that, of all the scalar quantifiers, "more than" should be mastered first, while "at most" should be last. Between them, Musolino's data and the findings of Experiment 3 confirm these predictions. More specifically, the combined evidence suggests the following picture. While 5-year-olds have serious trouble with superlative quantifiers, they are quite good with "more than". By the time they are 11, children are essentially perfect with "more than", still struggling with "at most", and fairly good with "at least" and "fewer than". This is precisely the pattern we expected to find.

7. Experiment 3

In the offline task of Experiment 1, adult participants produced correct responses on all trials,[6] so these data are in line with the null hypothesis that, for adults, all scalar quantifiers are equally easy (or hard). In our

6. More accurately: if we leave out of account the critical items, on which the standard theory and G&N's disagree, and two control items which unexpectedly prompted "at least" interpretations, 87% of the responses were correct; see Section 4.

third and last experiment we explored the possibility that there may be differences, after all, using an online design in which we first presented a statement, like "There are at least 2 A's", and then unveiled a "situation" consisting of one or more letters. Adult participants had to decide whether the statement was true of the situation, and we recorded reading times as well as decision times.

The reason why we separated trials into a reading stage and a decision stage is that we considered it likely that these stages correspond to cognitive processes which are distinct at least to some extent. In order to verify whether a sentence is true, it has to be interpreted first, and then the resulting interpretation must be confronted with the facts. Now, while the first part of this process may remain a relatively shallow affair, the second part requires a deeper understanding of the sentence in question, and since entailment properties and scalar type are "deeper" factors, our hypothesis was that their effects would be more pronounced at the decision stage than the reading stage.

Participants

32 students at the University of Cambridge, all native speakers of English, were paid for participating in this and five other experiments. The participants' ages ranged between 21;2 and 33;7; their mean age was 24;10. 23 of the participants were female.

Materials

All target statements were of the form "There are Q X's", where X was a letter (A or B) and Q was one of the following quantifiers: "more than 2", "at least 3", "fewer than 3", "at most 2", "exactly 3". Based on Musolino's acquisition data and the results of Experiment 2, we expected statements with "exactly" to be, if anything, easier than all others, so they were introduced as controls.

All situations were made up of 1, 2, 3 or 4 identical letters (A's or B's). For the control sentences ("exactly 3"), situations with 2, 3 and 4 letters were presented. For the other sentences, situations with 1, 2, 3 and 4 letters were presented. There were 38 trials in total. These were preceded by five practice trials, one for each quantifier. The order of the experimental trials was randomised for each participant.

Procedure

The experiment was run on a desktop computer using the E-Prime software package. On each trial, the target sentence was displayed first. Participants were instructed to press the space bar as soon as they had read and understood the sentence. Following a half-second delay, the sentence was then replaced by a situation. Participants had to decide as quickly as possible whether the sentence was true or false of the situation, and register their decision by pressing one of two keys. Reading times were recorded from sentence onset to the point at which the space bar was pressed: decision times were recorded from situation onset to the point at which the "yes" or "no" key was pressed.

Results

The average response times for all items are presented in Table 8. Four erroneous responses were removed from the decision times. A 2 $\times$ 2, Entailment (upward/downward) vs. Scalar Type (superlative/comparative), repeated measures ANOVA was run for the reading and the decision times. For the reading times, there were no significant effects or interac-

condition	*reading time (sd)*	*decision time (sd)*
Exactly	1580 (580)	1114 (220)
More than	1886 (911)	1271 (326)
At least	1921 (890)	1559 (606)
Fewer than	1940 (730)	1515 (431)
At most	1970 (719)	1982 (983)

Table 8: Mean response time in ms for Experiment 3.

tion between the two factors. With regard to the decision times, there was a main effect of Entailment ($F(1, 29) = 22.01$, $p < .001$, partial $\eta^2 = .43$) and a main effect of Scalar Type ($F(1, 29) = 17.78$, $p < .001$, partial $\eta^2 = .38$). The interaction between Entailment and Scalar Type was not significant ($F(1, 29) = 1.37$, n.s.).

Discussion

At the decision stage, these results accord with our main predictions. Besides the effect of Entailment, which was already attested in the literature, the main effect of Scalar Type supports the hypothesis that superlative quantifiers are harder to process than the comparative ones. That the pattern of reading times was different from what we observed at the decision

stage is in line with our conjecture that the experimental task involves two sub-tasks that tap into different cognitive processes: at the reading stage, entailment properties and the comparative/superlative distinction seem to have no effect whatsoever, which makes sense on the assumption that the interpretation processes that come into play at this stage are relatively shallow. By contrast, the decision stage demands a deeper understanding of the target statement, and it is here that the two semantic factors become critical.

8. Conclusion

As discussed in the introduction to this paper, it has long been thought that quantifiers are relatively simple devices, all assembled from the same puny inventory of building blocks, which makes the space of possible quantifiers fairly homogeneous. What we have called the standard view on quantifcation enforces homogeneity by assuming that the meanings of quantifying sentences like "All A are B", "Some A are B", "At most n A are B", etc. can always be analysed as set-theoretic statements about the set of A's and the set of B's: "All A are B" means that the set of A's is a subset of the set of B's, "Some A are B" means that the set of A's and the set of B's have a non-empty intersection, and so on. Even Chater and Oaksford (1999), who propose to jettison set theory in favour of probability theory, still seem to be adhering to the homogeneity view, arguing as they do that quantified statements should be analysed as probabilistic statements about the set of A's and the set of B's (e.g., "All A are B" is construed as $P(B|A) = 1$).

In psychology, the homogeneity assumption is perhaps nowhere as evident as in theories of syllogistic reasoning, which commonly presuppose that deduction is basically a matter of applying general-purpose rules to fairly simple representations, like logical formulae (Rips 1994) or mental models (Johnson-Laird and Byrne 1991), for example.

The most fundamental tenet of G&N's theory is that the meanings of quantifiers are not homogeneous: even if they appear to be closely related, the meanings of superlative quantifiers are essentially richer than those of their comparative counterparts, and not expressible in basic set theory. The experiments reported on in this paper confirm this view. Of course, this is not to say that they prove that G&N's theory is right, but they do show that G&N's theory does a better job than the standard account not only in linguistic but also in psychological terms. Furthermore, they add considerable support to the fundamental idea underlying

G&N's theory that superlative quantifiers have a richer semantics than standard ones.

On reflection, it shouldn't come as a surprise that quantifying expressions are a heterogeneous lot. After all, "every", "some", "most", etc., make up a tiny category of high-frequency words, and it would be quite remarkable if they were all alike. Indeed, it has been known for some time that, e.g., the meanings of "few" and "many" (Lappin 2000) and "any" (Kadmon and Landman 1993) appear to escape a straightforward analysis along the standard lines, and the same may hold for the differences between "all", "every", and "each", for example. If Geurts and Nouwen are right "at most" and "at least" can be added to this list.

Actually, it is not unthinkable that eventually all quantifiers will turn out to be special, but this is as it may be, for even if the homogeneity assumption had to be lifted for a handful of quantifiers only, it would be reason enough to reassess many if not most mainstream theories of quantification in logic, linguistics, and psychology.

References

Ariel, M. (2004). Most. *Language* 80: 658–706.

Barwise, J. and R. Cooper (1981). Generalized quantifiers and natural language. *Linguistics and philosophy* 4: 159–219.

Chater, N. and M. Oaksford (1999). The probability heuristics model of syllogistic reasoning. *Cognitive psychology* 38: 191–258.

Evans, J. S., S. E. Newstead, and R. M. Byrne (1993). *Human reasoning: the psychology of deduction*. Hove, East Sussex: Lawrence Erlbaum.

Geurts, B. (2003). Reasoning with quantifiers. *Cognition* 86: 223–251.

Geurts, B. (2006). Take "five": the meaning and use of a number word. In S. Vogeleer and L. Tasmowski (Eds.), *Non-definiteness and plurality*, pp. 311–329. Amsterdam/Philadelphia: John Benjamins.

Geurts, B. and R. Nouwen (2007). "At least" et al.: the semantics of scalar modifiers. *Language* 83: 533–559.

Geurts, B. and F. van der Slik (2005). Monotonicity and processing load. *Journal of semantics* 22: 97–117.

Gualmini, A. (2004). *The ups and downs of child language: experimental studies in children's knowledge of entailment relationships and polarity phenomena*. New York: Routledge.

Johnson-Laird, P. N. and R. M. Byrne (1991). *Deduction*. Hillsdale, New Jersey: Lawrence Erlbaum.

Just, M. A. and P. A. Carpenter (1971). Comprehension of negation with quantification. *Journal of verbal learning and verbal behavior* 10: 244–253.

Kadmon, N. and F. Landman (1993). Any. *Linguistics and philosophy* 15: 353 – 422.
Ladusaw, W. A. (1979). *Polarity sensitivity as inherent scope relations*. Ph.D. thesis, University of Texas at Austin.
Lappin, S. (2000). An intensional parametric semantics for vague quantifiers. *Linguistics and philosophy* 23: 599–620.
Moxey, L. M. and A. J. Sanford (1993). *Communicating quantities: a psychological perspective*. Hove/Hillsdale: Lawrence Erlbaum.
Musolino, J. (2004). The semantics and acquisition of number words: integrating linguistic and developmental perspectives. *Cognition* 93: 1–41.
Noveck, I. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition* 78: 165–188.
Pouscoulous, N., I. Noveck, G. Politzer, and A. Bastide (2007). A developmental investigation of processing costs in implicature production. *Language acquisition* 14: 347–376.
Rips, L. J. (1994). *The psychology of proof: deductive reasoning in human thinking*. Cambridge, Massachusetts: MIT Press.
van der Wouden, T. (1997). *Negative contexts: collocation, polarity and multiple negation*. London: Routledge.