To appear in: Journal of Semantics

Reasoning with 'some'

Bob van Tiel
Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS)

Ira Noveck
Institut des Sciences Cognitives – Marc Jeannerod

Mikhail Kissine
Université Libre de Bruxelles

Abstract

There has been substantial debate about the question of whether 'some' is normally interpreted as 'some but not all' when it is embedded under a quantifying expression. Experiments using a sentence-picture verification paradigm have been equivocal: while Geurts and Pouscoulous (2009) report that embedded upper-bounded construals of 'some' are almost non-existent, Potts and colleagues (2016) observed substantial rates of upper-bounded construals in at least some embedding environments. In this paper, we investigate how it is possible that these two superficially similar studies provide such diverging sets of results. We show that subtle features of the displays influence the frequency of embedded upper-bounded construals. We discuss the consequences of these findings for theories of upper-bounded construals and for experimental research in pragmatics in general.

1. Introduction

1.1. Wason's selection task

In his 1968 paper 'Reasoning about a rule', Wason describes an experiment aimed at testing whether people are competent logical reasoners, as supposed by Piaget's then popular theory of cognitive development (e.g., Inhelder & Piaget, 1959). In what eventually became known as the Wason Selection Task, participants are presented with four cards showing, e.g., A, D, 4, and 7, along with the information that each card has a letter on one side and a number on the other. Participants are then provided with an arbitrary conditional rule, such as 'If there is a vowel on one side, then there is an even number on the other.' Their task is to turn over those cards, and only those cards, that can help determine whether the rule is true or false.

From a logical point of view, the correct answer would be to turn over the cards showing A and 7, because these are the cards that can potentially falsify the conditional rule, e.g., if the other sides of those cards showed 5 and E. However, participants rarely select these cards. Instead, the most popular option is to turn over either the cards showing A and 4, or just the card showing A. In other words, participants are biased towards confirming rather than falsifying the conditional rule. This confirmation bias led Wason to conclude that adults have systematic difficulties with logical reasoning, thus speaking against Piaget's theory of cognitive development.

In the years that followed, however, it became apparent that the results of Wason's original study were contingent on the type of materials that were used, and that rules with a different content could greatly increase the proportion of logically correct responses (e.g., Johnson-Laird, Legrenzi, & Legrenzi 1972, Wason & Shapiro 1971). To illustrate, Griggs and Cox (1982) asked participants to imagine being a police officer entering a bar to *enforce* the following conditional rule 'If a person is drinking beer, then the person must be over 19 years of age'. Patrons were represented by cards showing what they were drinking on one side and their age on the other side. Participants were thus presented with four cards reading 'drinking a beer', 'drinking a coke', '22 years of age', and '16 years of age', and they were asked to select the card or cards that had to be turned over in order to find violators of the rule. In this scenario, most participants made the logically correct decision of turning over the cards reading 'drinking a beer' and '16 years of age'.

In the wake of these conflicting findings, numerous researchers then attempted to tease apart the relevant factors that determine how participants behave in the

Wason Selection Task (e.g., Cheng & Holyoak, 1985, Cosmides & Tooby, 1992, Noveck & O'Brien, 1996, Oaksford & Chater, 2007, Sperber, Cara, & Girotto, 1995, Stenning & van Lambalgen, 2004). What this line of research has shown is that it is a mistake to assume that the question of whether people are competent logical reasoners can be answered with a simple yes or no. Rather, the extent of people's logical abilities is heavily dependent on methodological features of the task itself.

1.2. Background on 'scalar implicature'

As will become clear, the debate about the Wason Selection Task serves as an instructive analogy to the current debate in pragmatics about so-called 'embedded implicatures'. This debate concerns the interpretation of scalar expressions such as 'some' and 'or' when they are embedded, e.g., under other quantifying expressions, such as 'every' and 'exactly one'. Before turning to these controversial cases, first consider the intensively studied unembedded case, as illustrated in (1).

(1) Some of the students passed the exam.

The speaker of this sentence may imply that not all of the students passed the exam. On this interpretation, the scalar expression 'some' receives an upper-bounded construal (UBC) and thus comes to exclude 'all'. To explain this UBC, one can consider two types of mechanisms: *Gricean reasoning* and *truth-conditional narrowing*.

To illustrate the notion of Gricean reasoning, suppose someone utters (1). The hearer may reason as follows: the speaker could have been more informative by saying 'All of the students passed the exam'. Why didn't she? Presumably because she does not believe it is true that all of the students passed the exam. This *ignorance inference* can be strengthened if it is likely that the speaker knows whether or not all of the students passed the exam. If this *competence assumption* holds, it follows that the speaker believes it is false that all of the students passed the exam (e.g., Gazdar, 1979, Geurts, 2010, Horn, 1972, Soames, 1982).

In the case of truth-conditional narrowing, the semantic meaning of the scalar expression is restricted to exclude its upper bound. In this case, then, the UBC becomes part of the compositional meaning of the uttered sentence rather than being a post-compositional pragmatic inference. A separate question, which is orthogonal to the immediate purpose of this paper, concerns the nature of truth-conditional narrowing. Various proposals have been made to account for the process that underlies truth-conditional narrowing. Here, we briefly describe three such proposals.

i. Scalar expressions such as ‘some’ might be lexically ambiguous between a lower-bounded ‘some and possibly all’ meaning and a two-sided ‘some but not all’ meaning (e.g., Ariel, 2015, Levinson, 2000, Storto & Tanenhaus, 2005).

ii. Truth-conditional narrowing might be due to the presence of a covert syntactic operator, whose meaning is essentially that of overt ‘only’. Appending this operator to (1) leads to an interpretation that can be paraphrased as ‘Only some students passed the exam’, thus accounting for the UBC (e.g., Chierchia, 2004, 2006, Chierchia, Fox, & Spector, 2012).

iii. Truth-conditional narrowing might have a pragmatic source. Just like the meaning of ‘drink’ in ‘Can I offer you a drink’ can be narrowed to mean ‘drink alcohol’, so too, according to this pragmatic view, the meaning of ‘some’ can be narrowed to mean ‘some but not all’ given the right contextual circumstances (e.g., Geurts, 2009, Noveck & Sperber, 2007, Nunberg, 1978). A particularly elegant and precise model of pragmatic narrowing is given by Potts, Lassiter, Levy, and Frank (2016), who argue that rational principles such as the assumptions of informativeness and conciseness also operate at the subsentential level.

The differences and the relative merits of these three proposals are unimportant for the moment. What is important is that, in the case of truth-conditional narrowing, the semantic content of the scalar expression is modified to exclude the upper bound, whereas, in the case of Gricean reasoning, ‘some’ retains its logical meaning as ‘some and possibly all’ and the upper bound is communicated as a pragmatic inference. In other words, in the case of narrowing, the upper bound is *expressed*; in the case of Gricean reasoning, it is merely *communicated*, assuming that what is expressed is a subset of what is communicated (Horn, 2004).

There is general agreement that both Gricean reasoning and truth-conditional narrowing variously underlie UBCs. What current theories disagree about is the division of labour between these two mechanisms. On the one hand, *pragmatic* theories hold that Gricean reasoning is the norm and that truth-conditional narrowing is restricted to marked cases involving prosodic emphasis on ‘some’ or a salient contrast with ‘all’, as illustrated by the following sentence (cf. Horn 2006, p. 26, see also, e.g., Geurts 2010, Geurts & van Tiel 2013, Noveck & Sperber 2007):

(2) If SOME of my friends come to the party, I’ll be happy—but if ALL of them do, I’ll be in trouble.

According to pragmatic theories, someone who hears this sentence being uttered has to reinterpret the scalar expression ‘some’, e.g., upon encountering its stronger scalemate ‘all’ in order to make sense of it. Such theorists would further agree that, while there is no cut-and-dried method for determining under what circumstances this type of reinterpretation occurs, it should be uncommon and marked by prosodic or contextual means.

On the other hand, *conventionalist* theories hold that the scope of Gricean reasoning is limited and that UBCs are normally the result of truth-conditional narrowing (e.g., Chierchia 2004, 2006, Chierchia et al. 2012, Levinson 2000, Potts et al. 2016). To illustrate, Chierchia and colleagues (2012) invoke Gricean reasoning to explain that, in the absence of speaker competence, uttering (1) merely implies that the speaker does not believe that all of the students passed, while ascribing to truth-conditional narrowing the stronger inference that the speaker believes it is false that all of the students passed.

1.3. ‘Embedded implicatures’

In order to decide between pragmatic and conventionalist accounts of UBCs, both theorists and experimentalists have turned their attention to the interpretation of ‘some’ when it is embedded under a quantifying expression, as in (3a-c) below.

(3) a. Every student passed some of the exams.
b. Exactly one student passed some of the exams.
c. None of the students passed some of the exams.

An utterance of (3a) may imply that not all students passed all of the exams, and an utterance of (3b) may imply that exactly one student passed some but not all of the exams, with all other students passing none of them. These *global* UBCs can be explained both in terms of Gricean reasoning and truth-conditional narrowing.

In at least some cases, an utterance of (3a) may also imply that no student passed all of the exams, an utterance of (3b) may also imply that exactly one student passed some but not all of the exams, with all other students passing either none or all of them, and an utterance of (3c) may imply that all students passed either none or all of the exams. These *embedded* UBCs can be paraphrased by replacing ‘some’ in the sentences above with ‘some but not all’ and can be brought out, e.g., by explicitly contrasting ‘some’ with ‘all’:

(4) Exactly one student passed SOME of the exams—the other students passed ALL of them.

Embedded UBCs, however, cannot be derived by means of Gricean reasoning and hence indicate the presence of truth-conditional narrowing. The point in question, then, is how frequently hearers derive these embedded UBCs.

Pragmatic theories predict that embedded UBCs should be rare and restricted to situations in which the contrast between 'some' and 'all' is particularly salient (e.g., Geurts, 2010, Geurts & van Tiel, 2013). Some conventionalist theories predict the converse, i.e., that embedded UBCs should be commonplace across the board. Other conventionalist theories have proposed auxiliary assumptions to constrain the prevalence of truth-conditional narrowing.

One such constraint holds that UBCs are restricted to situations in which they lead to a logically stronger interpretation—i.e., an interpretation that is true in a more restricted set of situations. This constraint predicts that only (3a) should normally license an embedded UBC (Chierchia, 2006, Chierchia et al., 2012). Another constraint holds that UBCs are restricted to situations in which they do not lead to a logically weaker interpretation—i.e., an interpretation that is true in a less restricted set of situations. This constraint predicts that both (3a) and (3b) should normally license an embedded UBC (Chemla & Spector, 2011).

The predictions of the conventionalist approach of Potts and colleagues (2016) are less categorical in this respect. The central tenet behind their *lexical uncertainty model* is that hearers are uncertain about the meaning of the words that the speaker uses. To illustrate, consider the ubiquitous 'some'. The lexical meaning of 'some' can be paraphrased as 'some and possibly all'. However, according to the lexical uncertainty model, someone who hears an utterance containing 'some' also considers it possible that the speaker had a more specific meaning in mind, such as 'some but not all' or even 'all'. In particular, the hearer determines, for each possible meaning of the words the speaker used, which sentences the speaker would have produced in which situations. Based on this information, the hearer constructs a conditional probability distribution from utterances to interpretations.

A linking hypothesis is needed to map these conditional probabilities to rates of embedded UBCs. One possible linking hypothesis would be to normalise the conditional probabilities to the interval $[0, 1]$. Based on this linking hypothesis, using the same parameters as Potts and colleagues use, the lexical uncertainty model predicts that the rates of embedded UBCs fluctuate across embedding environments, but are generally in excess of 50%.[1]

[1] The predictions of the lexical uncertainty model depend on a number of parameters, including the set of alternatives and the presumed "greediness" with which the hearer extracts information from the behaviour of the speaker. Changing these parameters from the values assigned to them by

To sum up, pragmatic theories predict that ‘some’ in the sentences in (3) should normally not be interpreted as ‘some but not all’. By contrast, most conventionalist theories predict that at least one and possibly all of the sentences in (3) should give rise to substantial rates of embedded UBCs.

The interpretation of sentences such as (3) has been tested in a number of experimental studies using a range of different tasks (e.g., Benz & Gotzner, 2017, Chemla & Spector, 2011, Clifton & Dube, 2010, Franke, Schlotterbeck, & Augurzky, 2017, Geurts & Pouscoulous, 2009, Potts et al., 2016, Tian, Breheny, & van Tiel, 2012). There has been a substantial amount of equivocation about the results of at least some of these tasks. In particular, it has been argued that tasks that ask for graded truth judgements or preferred interpretations might measure typicality structure rather than the prevalence of UBCs (van Tiel, 2014).

One of the tasks that does not face this interpretational ambiguity and that can boast a long lineage in experimental psychology is the *sentence-picture verification task* (e.g., Abrams, Chiarello, Cress, Green, & Ellett, 1978, Clark & Chase, 1972). In this type of task, participants are presented with a sentence and a display and have to indicate if the sentence is an accurate description of the corresponding display. Sentence-picture verification tasks are common in psychological research and have been validated in numerous studies on *un*embedded UBCs (e.g., Barner, Brooks, & Bale, 2011, Degen & Tanenhaus, 2011, Papafragou & Musolino, 2003, Pouscoulous, Noveck, Politzer, & Bastide, 2007, van Tiel & Schaeken, 2017). An example item, based on the items used in the study of van Tiel and Schaeken, is provided in Fig. 1. In this item, the sentence is literally true but false if ‘some’ is interpreted as ‘some but not all’. Hence, the proportion of ‘false’ responses—which tends to range between 25% and 80% in these studies (cf. Dieussaert, Verkerk, Gillard, & Schaeken 2011, p. 2362; Geurts, 2010, p. 99)—provides a measure of the frequency with which the corresponding UBC is derived.

Two studies, to which we will turn shortly, have made use of a sentence-picture verification task to determine the frequency of embedded UBCs, and have reported markedly different results: whereas Geurts and Pouscoulous (2009) found that embedded UBCs are extremely rare, Potts and colleagues (2016) observed substantial rates of UBCs in at least some embedding environments.

These conflicting findings are reminiscent of the observation summarised in the introduction, i.e., that participants in the Wason Selection Task provide substantially more logically correct responses when the task is concrete and normative

Potts and colleagues has important consequences for the predicted rates of embedded UBCs. We return to this issue in Section 4.2.

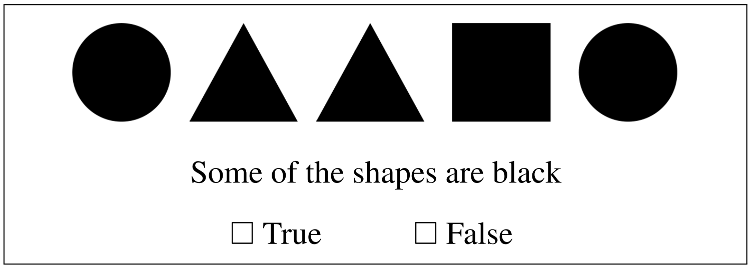


Figure 1: Example item for testing the frequency of unembedded UBCs, based on the study of van Tiel and Schaeken (2016).

(Griggs & Cox, 1982) than when it is abstract and descriptive (Wason, 1968). Researchers on the Wason Selection Task have generally accepted these conflicting results and sought to come to an understanding of the factors that are responsible for the discrepancy (e.g., Noveck & O'Brien, 1996, Stenning & van Lambalgen, 2004). This contrasts with the way researchers on 'embedded implicatures' have reacted, which is to be unusually selective in assuming that only one set of results can provide a correct estimation of the frequency of embedded UBCs. The upshot is that individual researchers view contradictory findings as a result of methodological inadequacies in one direction or the other. In this way, a number of authors have argued—non-productively in this context—that the study of Geurts and Pouscoulous is significantly flawed (e.g., Benz & Gotzner, 2014; Chemla & Spector, 2011, p. 364–367; Clifton & Dube, 2010, p. 3–4; Potts et al., 2016, p. 779).

The goal of this paper is to adopt the approach from the Selection Task literature and make sense of the extant data rather than simply choosing the data set that affirms one account while disparaging others'—a move that brings to mind the confirmation bias that the selection task and another Wason task (the 2–4–6 task, cf. Wason, 1960) made famous. Before turning towards the goal of explaining the conflicting data, however, we first consider the studies of Geurts and Pouscoulous and Potts and colleagues in more detail.

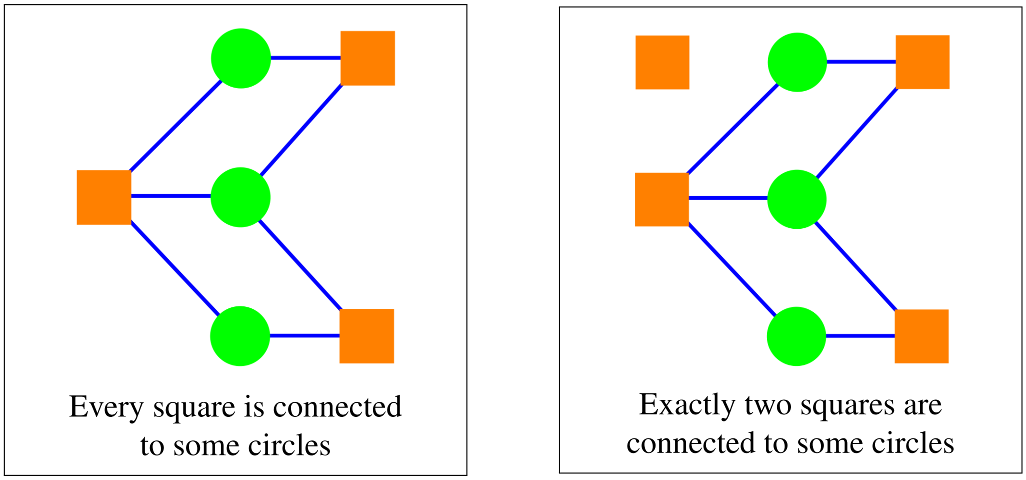


Figure 2: Example critical items from Geurts and Pouscoulous (2009: Exp. 3).

2. *Previous work*

2.1. Geurts and Pouscoulous (2009)

Geurts and Pouscoulous (2009, Exp. 3) were the first to experimentally investigate the interpretation of ‘some’ in embedded environments. They presented 26 Dutch students in the humanities with booklets showing, on each page, a sentence and a display consisting of coloured squares and circles on a partially filled 3×3 grid. Participants had to indicate whether the sentences that described these displays were true or false. The experiment consisted of 6 critical items that were interspersed with 37 filler items.

The 6 critical items showed sentences with ‘enkele’, which is roughly equivalent to English ‘some’, embedded under a quantifying expression. Geurts and Pouscoulous tested the Dutch equivalents of the following quantifying expressions: ‘every’, ‘more than one’, ‘exactly two’ (which was tested twice), ‘not all’, and ‘not more than one’. The corresponding displays either verified the sentence on its literal interpretation and on its interpretation with a global UBC, but falsified it on its interpretation with an embedded UBC, or, conversely, falsified the sentence on its literal interpretation and on its interpretation with a global UBC, but verified it on its interpretation with an embedded UBC. The truth judgements that participants gave were thus indicative of whether or not they had derived the corresponding embedded UBC. Two example items are shown in Fig. 2.

Geurts and Pouscoulous (2009)

Quantifier	Display	UBC	%
‘every’	SSA	F	0
‘exactly two’	NNSA	F	0
‘exactly two’	NSSA	T	0
‘not more than one’	NNSA	T	4

Potts and colleagues (2016)

Quantifier	Display	UBC	%
‘every’	SAA	F	8
‘exactly one’	NNA	F	27
‘exactly one’	SAA	T	51
‘no’	NAA	T	43

Table 1: Percentages of responses compatible with an embedded UBC of ‘some’ for a selection of the critical conditions in the two studies. UBC: Truth value if an embedded UBC is derived.

In what follows, we use the following notational conventions to describe these items: the letter A refers to a square that is connected to all of the circles, S to a square that is connected to some but not all of the circles, and N to a square that is connected to none of the circles. The corresponding target sentences will be denoted by the quantifying expressions under which ‘some’ is embedded. In summary, then, the items in Fig. 2 will be referred to as ‘every’-SSA and ‘exactly two’-NSSA, respectively.

In the ‘every’-SSA condition, shown in the left panel of Fig. 2, the sentence is true if interpreted literally or if a global UBC is derived, but false if it is interpreted with an embedded UBC, since one of the squares is connected to all of the circles. Conversely, in the ‘exactly two’-NSSA condition, shown in the right panel of Fig. 2, the sentence is false if interpreted literally or if a global UBC is derived, but true if interpreted with an embedded UBC, since exactly two of the squares are connected to some but not all of the circles.

A selection of the results is presented in Table 1. On the whole, Geurts and Pouscoulous observed negligible rates of responses that are indicative of embedded UBCs, which did not exceed 4% in any of the conditions. In this respect, their results differ markedly from what Potts and colleagues (2016) observed in a superficially similar experiment, to which we now turn.

2.2. Potts, Lassiter, Levy, and Frank (2016)

Potts and colleagues conducted their experiment online on Amazon’s Mechanical Turk. 800 participants saw, on each trial, a sentence and a display showing three basketball players, labelled A, B, and C. Each basketball player was flanked by two (possibly empty) piles of basketballs, which were labelled ‘baskets’ and ‘misses’.

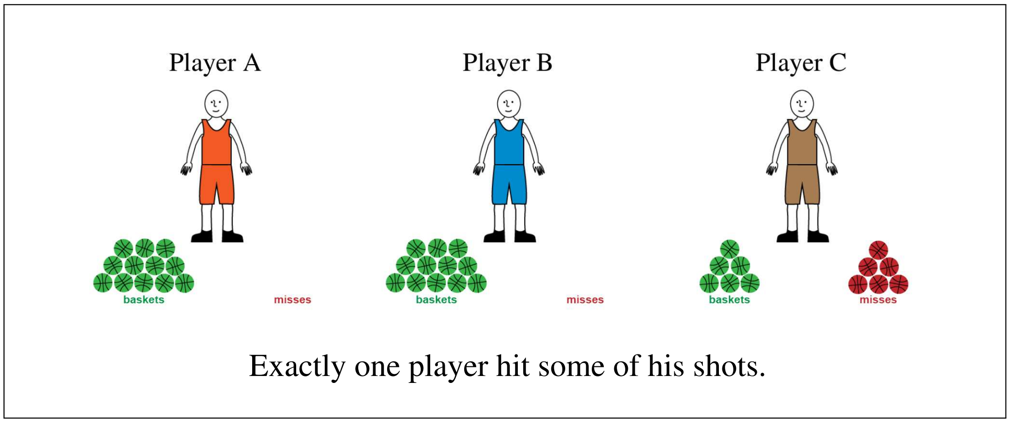


Figure 3: Example critical item from Potts et al. (2016).

The total number of balls in these two piles was always 12. The displays represented the outcome of a basketball free-throw shooting competition.

According to the instructions, the corresponding sentences had been generated by an automated sportscasting system. The cover story for participants was that they had to indicate whether the sentences were true or false so that the automated sportscasting system could be trained further. The experiment consisted of 3 critical items, 6 control items, and 23 filler items, and was preceded by a practice phase consisting of 3 filler items.

The 3 critical items showed sentences with ‘some’ embedded under ‘every’, ‘exactly one’, or ‘no’. The corresponding displays consisted of three types of players: those who made all 12 shots (which we refer to as A), those who made 6 shots and missed 6 shots (S), and those who missed all 12 shots (N). Participants were presented with displays showing random distributions of these A, S, and N cases. An exemplary critical item is shown in Fig. 3. Control items were identical to critical items, except that ‘some’ was replaced by either ‘all’ or ‘none’.

Filler items notably included two test sentences with the quantifying expression ‘most’ in an unembedded position. As one of our reviewers pointed out, these filler items might have caused participants to derive an unembedded UBC, thus interpreting ‘most’ as ‘most but not all’. Ostensibly, none of the filler items in the study of Geurts and Pouscoulous could give rise to unembedded UBCs.

Some of the critical items were of particular interest because the truth value of the sentence depended on whether or not an embedded UBC was derived. In this paper, we focus on four of these items: ‘every’-SAA, ‘exactly one’-NNA, ‘exactly one'-SAA (shown in Fig. 3), and 'no'-NAA. These critical items were most similar to the critical items tested by Geurts and Pouscoulous and were associated with the highest rates of embedded UBCs. The rates of embedded UBC responses for these critical items are shown in Table 1.

In the case of 'every'-SAA, the rate of embedded UBC responses was still quite marginal. The 'exactly one'-SAA and 'no'-NAA cases, however, provide evidence that participants frequently derived the corresponding embedded UBCs.

2.3. Summary and outlook

In summary, then, the results presented by Geurts and Pouscoulous strongly support the pragmatic view that 'some' is normally not interpreted as 'some but not all' when it is embedded under a quantifying expression. By contrast, the results presented by Potts and colleagues indicate that such embedded UBCs are quite frequent—at least for 'exactly one'-SAA and 'no'-NAA. Only the latter set of findings is supportive of a conventionalist approach.

At this point, the two studies differ in a large number of respects, thus precluding a straightforward comparison. This is why we will first repeat the studies in a more uniform experimental setting to determine the robustness of both sets of results and to obtain a solid basis of comparison, before exploring the effects of methodological features that might influence the frequency of embedded UBCs.

Hence, in Exp. 1, we repeated the two studies using the same pool of participants (Mechanical Turk users), conducting the experiment in the same language (American English), with the same number and types of fillers, the same quantifying expressions ('every', 'exactly one', and 'no'), the same scalar expression ('some of the'), the same instructions, and the same general procedure. The two conditions in Exp. 1 thus differ in only two respects: the nature of the displays and the distribution of A, S, and N cases in the critical items.

Experiment 1: Levelling the playing field

Participants

240 participants were drafted on Amazon's Mechanical Turk. Half of them participated in the pseudo-replication of Geurts and Pouscoulous (hence, the *squares-and-circles task*); the other half in the pseudo-replication of Potts and colleagues (hence, the *basketball task*). 31 participants were removed either because their native language was not English or because they had participated in a similar

Squares-and-circles

Quantifier	Display	Other	UBC
‘every’	SSA	T	F
‘exactly one’	NNNA	T	F
‘exactly one’	NNSA	F	T
‘no’	NNNA	F	T

Basketball

Quantifier	Display	Other	UBC
‘every’	SAA	T	F
	(not tested)		
‘exactly one’	SAA	F	T
‘no’	NAA	F	T

Table 2: Critical items tested in Exp. 1. Other: Truth value if the sentence is interpreted literally or if a global UBC is derived. UBC: Truth value if an embedded UBC is derived.

experiment before. The mean age of the remaining participants was 34 (standard deviation: 10). 95 of the participants were female.

Materials

The experiment contained 4 (squares-and-circles) or 3 (basketball) critical items, as shown in Table 2. The critical sentences contained the partitive ‘some of the’ embedded under a quantifying expression. Note that Geurts and Pouscoulous tested the bare form ‘enkele’, which is roughly equivalent to English ‘some’, instead of the corresponding partitive. Three embedding quantifying expressions were tested: ‘every’, ‘exactly one’, and ‘no’. Note that we tested ‘exactly one’ twice in the squares-and-circles task and once in the basketball task so that the two tasks were as similar as possible to the original studies.

These critical items were interspersed with 13 filler items. The fillers for the basketball task were a sample of the fillers used in the original study of Potts and colleagues. These fillers included one item with ‘most’ in unembedded position. The fillers were ‘translated’ for the squares-and-circles task to fit the corresponding displays. Three filler items and their ‘translations’ are provided in (5) and (6). A complete list of the filler items in both tasks is provided in the Appendix.

(5) **Basketball**

a. Most of the players missed shots.
b. Player A placed second.
c. We have a clear loser.

(6) **Squares-and-circles**

a. Most of the squares are connected to circles.
b. The red circle is the second most connected shape.

c. One of the circles has the fewest connections.

The displays in the squares-and-circles task always consisted of coloured squares and circles on a partially filled 3×3 grid. The critical displays were analogous to the ones used in the original study of Geurts and Pouscoulous, except for the 'no'-NNNA case, which was not tested in their study. The displays in the basketball task were identical to the ones used in the original study of Potts and colleagues. Critical items from both tasks are shown in Figs. 4 and 5.

The truth value of the filler items was pseudo-randomised across five lists. That is, for each filler sentence, one display was created that verified the sentence and one display that falsified it. For each list, one display was randomly selected for each filler sentence, making sure that the list contained an approximately equal number of verifying and falsifying displays. Moreover, in the basketball task, these lists also varied the order of the A, S, and N cases in the critical items. For example, whereas the first list would test 'no'-NAA, the second list tested 'no'-ANA, and the third list tested 'no'-AAN. The order of presentation was randomised for each participant.

Procedure

The instructions for participants in the basketball task were copied from the original experiment of Potts and colleagues and went as follows:

> We are trying to train an automated sportscasting system to generate color commentary on simple sports competitions. We'd like you to make judgments about the comments it generates. We will use these ratings to train our system further.
>
> On each page, we will show you the outcome from a basketball free-throw shooting competition between three players, labeled A, B, and C for convenience. We will give you a comment the system has generated, and your job is to decide whether this comment is true or false.

These instructions were suitably adapted for the squares-and-circles task in order to fit the corresponding displays. For example, 'an automated sportscasting system' was replaced by 'an automated system', and 'sports competitions' was replaced by 'simple displays'.

Participants were further instructed to register their truth judgements by pressing either '1' (true) or '0' (false) on their keyboard. Items were separated with a fixation cross, which appeared on screen for 1 second.

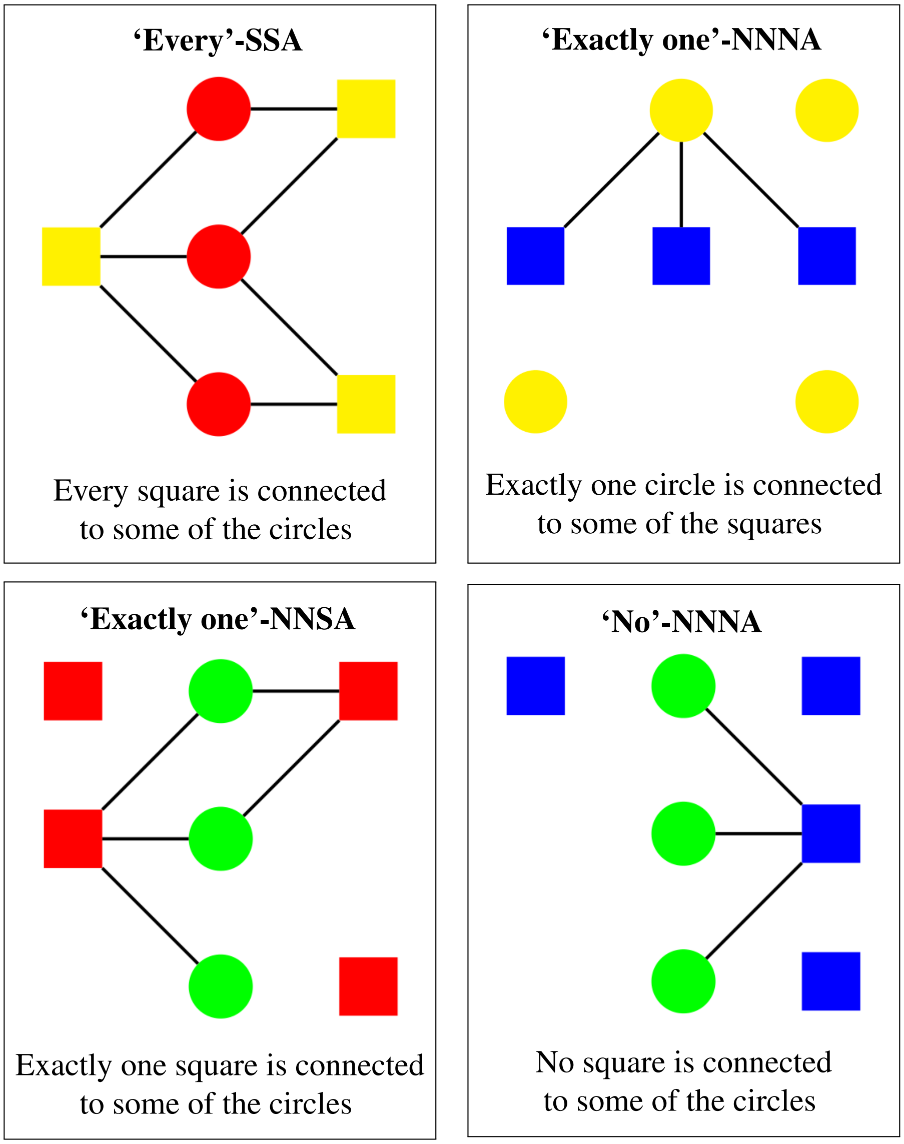

‘Every’-SSA
Every square is connected to some of the circles
‘Exactly one’-NNNA
Exactly one circle is connected to some of the squares
‘Exactly one’-NNSA
Exactly one square is connected to some of the circles
‘No’-NNNA
No square is connected to some of the circles

Figure 4: Critical items tested in the squares-and-circles task (Exp. 1).

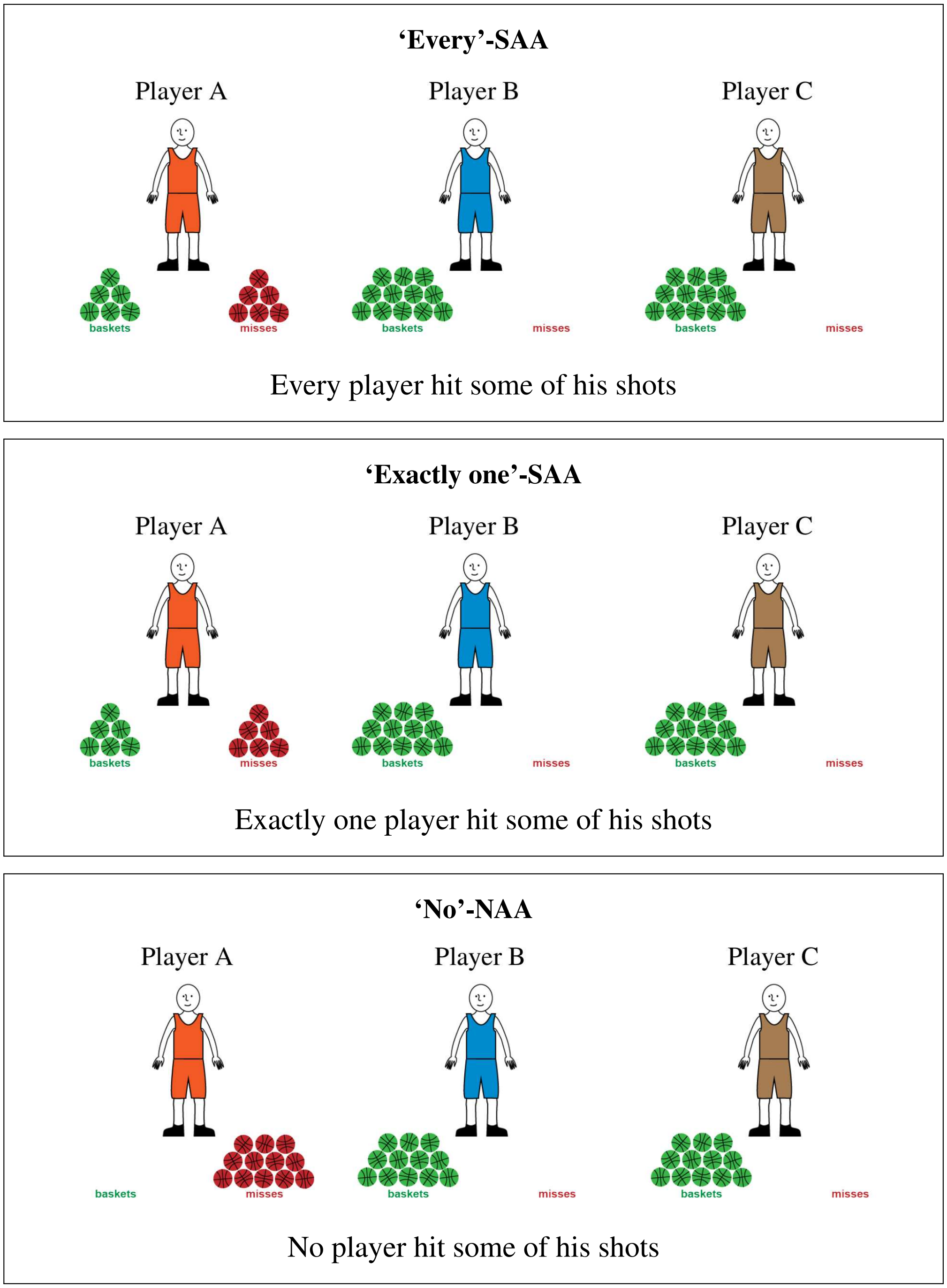


Figure 5: Example critical items tested in the basketball task (Exp. 1).

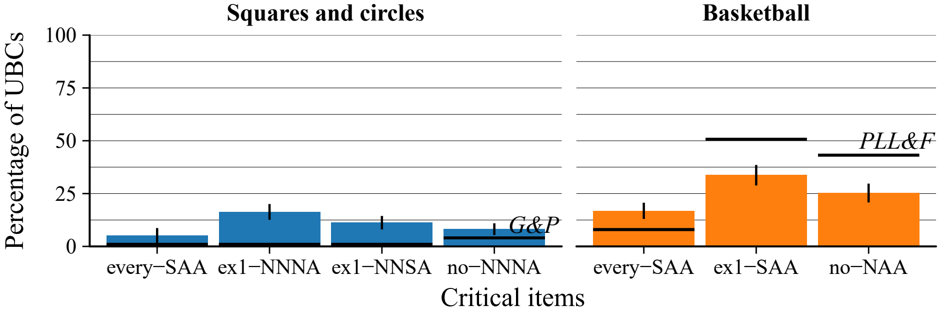


Figure 6: Percentages of embedded UBC responses in Exp. 1. Horizontal lines represent the percentages of embedded UBC responses in the original studies of Geurts and Pouscoulous (G&P) and Potts and colleages (PLL&F). (Note that 'no'-NNNA is paired with the results of the similarly negative 'not more than one'-NNSA.) Error bars represent standard errors.

Results

6 participants were removed from the analysis for making mistakes in more than 20% of the filler items. This leaves 98 (squares-and-circles) and 105 (basketball) participants. The mean error rate on filler items of the remaining participants was 5.5%. The percentages of embedded UBC responses are shown in Fig. 6. In the following analysis, we ignore the 'exactly one'-NNNA item, which was only tested in the squares-and-circles task. Moreover, we group together the critical items based on the quantifying expression under which 'some' occurs. See Section 3 for an experiment in which we test whether the structure of the displays affects the probability with which 'some' is interpreted as 'some but not all'.

First of all, the results confirm that there is a substantial difference in the rates of embedded UBCs between the two tasks: 10.2% in the squares-and-circles task and 25.3% in the basketball task. To analyse whether this difference was statistically significant, we constructed a mixed effects logistic regression model predicting response (UBC or other) on the basis of task (squares-and-circles or basketball) with random intercepts for participants and items. This analysis confirmed that the difference was significant (β = -1.66, *SE* = 0.34, *Z* = -4.83, $p < .001$). Looking in more detail, using chi-square tests of proportions, we observe that the difference was significant for all critical items (all p's $< .007$). In summary, then, the results of Exp. 1 confirm that the rates of embedded UBCs are sensitive to the materials used in the task.

At the same time, however, the results of Exp. 1 indicate that the difference between the two tasks is not as big as the difference between the two original studies. On the one hand, the squares-and-circles task was associated with rates of embedded UBCs that consistently exceeded the 4% observed by Geurts and Pouscoulous. A possible explanation for this difference is that Exp. 1 tested the partitive ‘some of the’ instead of the bare form ‘enkele’ (= ‘some’) that was tested by Geurts and Pouscoulous. It has been observed that partitives lead to higher rates of UBCs than bare forms (e.g., Banga, Heutinck, Berends, & Hendriks, 2009).

In line with this explanation, Geurts and Pouscoulous, in their Exp. 4, report the results of a slightly different sentence-picture verification task, in which participants could also provide ‘could be either’ as a possible response. In that experiment, which tested 22 students of University College London, the critical items focused on the partitive ‘some of the’. The rates of embedded UBC responses in that experiment were more in line with the results of our squares-and-circles task: 5% for ‘every’-SSA, 5% for ‘exactly two’-NNSA, and 9% for ‘exactly two’-NSSA.

On the other hand, the basketball task was associated with rates of UBC responses that were markedly lower than what Potts and colleagues found, especially for ‘exactly one’-SAA (36% vs. 51%) and ‘no’-NAA (25% vs. 43%). Using the data from Potts and colleagues, we observed that these differences were either significant or marginally significant ($\chi^2 = 3.62$, $p = .06$ for ‘exactly one’-SAA and $\chi^2 = 7.06$, $p = .008$ for ‘no’-NAA). In order to arrive at a satisfactory explanation of this discrepancy, we investigated three differences between our basketball task and the original study of Potts and colleagues.

First, in the original study, the experiment was preceded by a short practice phase, whereas our study had none. Second, the original study included control items that were structurally similar to the target items but had the scalar items ‘none’ and ‘all’ instead of ‘some’. Such control items could serve to render salient the contrast between ‘some’ and the other quantifying expressions. Third, the original study was twice as long as our basketball task. We intuited that UBCs might emerge over the course of the task.

To test our intuition that UBCs emerge over the course of the task, we constructed a mixed effects logistic regression model predicting response (UBC or other) on the basis of trial number, with random intercepts for participants, items, and tasks (squares-and-circles or basketball) using the data from Exp. 1. We observed a significant effect of trial number ($\beta = 0.05$, $SE = 0.02$, $Z = 2.02$, $p = .04$), providing tentative evidence for our intuition that UBCs emerge over the course of the task.[2]

In what follows, we refer to the three parameters that might have been responsible for the discrepancy between the results of our basketball task and the results of the original study of Potts and colleagues as P(ractice), C(ontrol), and L(ength). Each of these parameters has two settings, i.e., present (+) or absent (–) in the case of P and C, and long (+) or short (–) in the case of L. To reiterate, the basketball task in Exp. 1 was P–C–L–.

In the next section, we present the results of an experiment in which we manipulated these three parameters. Although we did not test all eight possible parameter settings, the results of the four combinations that we tested, i.e., P+C+L+ (which is identical to the parameter settings in the original study of Potts and colleagues), P–C+L+, P+C–L+, and P+C–L–, along with the results of the basketball task used in Exp. 1 (i.e., P–C–L–), provide sufficient grounds for drawing conclusions about the respective roles of these parameters.

Experiment 2: Practice, control, length

160 participants were drafted on Mechanical Turk. Participants were equally distributed across the four conditions (P+C+L+, P–C+L+, P+C–L+, and P+C–L–). 28 participants were removed from the analysis either because their native language was not English or because they had participated in similar experiments before. The mean age of the remaining participants was 34 (standard deviation: 11). 46 of the participants were female.

Materials

The materials were the same as in the basketball task of Exp. 1. That is, there were always 3 target items: ‘every’-SAA, ‘exactly one’-SAA, and ‘no’-NAA. These target items were interspersed with either 13 (L–) or 29 (L+) filler items, which were taken from or based on the original study of Potts and colleagues. A complete list of the filler items is provided in the Appendix. In the C+ condition, 6 filler items were replaced with control items, which were structurally similar to target items, except that ‘some’ was replaced with either ‘all’ or ‘none’.

[2] We conducted the same analysis on the data from Potts and colleagues, where the effect of trial number on the probability of embedded UBC responses went in the same direction but was not significant ($\beta = 0.02$, $SE = 0.02$, $Z = 1.51$, $p = .13$).

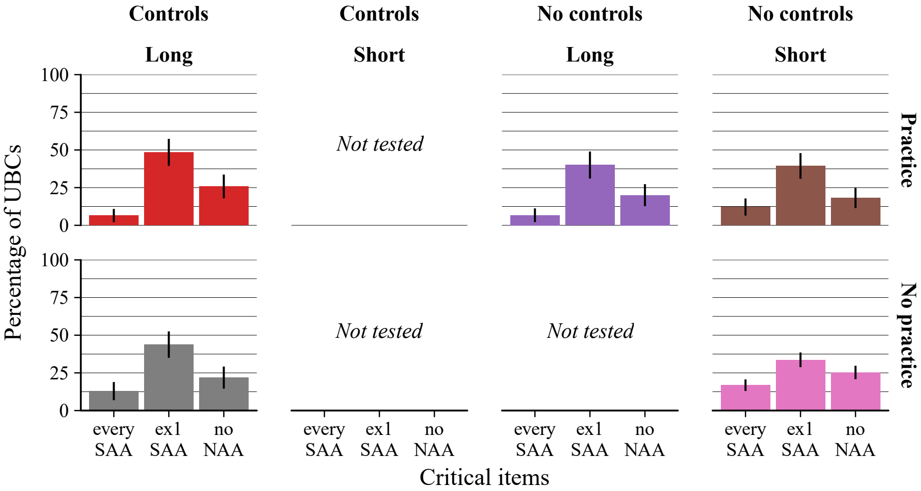


Figure 7: Percentage of embedded UBC responses in the basketball task of Exp. 1 (P–C–L–) and in the four conditions of Exp. 2 (P+C+L+, P–C+L+, P+C–L+, P+C–L–). We did not test the remaining 3 parameter settings (P+C+L–, P–C+L–, P–C+L+). Error bars represent standard errors.

Procedure

The procedure was the same as in the basketball task of Exp. 1. In the P+ condition, however, the experiment was preceded by a practice phase consisting of 3 filler items. As in the original study of Potts and colleagues, the practice phase started with the message 'We're going to do some practice now' and ended with the message 'OK, let's get started for real'. The practice phase did not include feedback on whether or not the responses were correct.

Results

9 participants were removed from the analysis for making mistakes in more than 20% of the filler items. The mean error rate of the remaining participants was 4.5%. The percentages of UBC responses in each of the conditions are shown in Fig. 7, along with the percentages of UBC responses observed in the basketball task of Exp. 1 (i.e., P–C–L–).

In order to determine whether the rates of embedded UBCs differed across conditions, we constructed a mixed effects logistic regression model predicting responses (UBC or other) on the basis of task (P+C+L+, P+C–L+, P–C+L+, P–C–

L+, or P–C–L–) and condition (‘every’-SAA, ‘exactly one’-SAA, or ‘no’-NAA), with random intercepts for participants.[3] We compared the fit of this model with that of a similar model that did not include task as a fixed factor, by means of a chi-square test. This test indicated that including task as a fixed factor did not significantly improve the model fit ($\chi^2(4) < 1$).

In order to determine whether the rates of embedded UBCs in the tasks tested in Exp. 2 (i.e., P+C+L+, P+C–L+, P–C+L+, and P–C–L+) differed from the corresponding rates in the study of Potts and colleagues, we conducted chi-square tests of proportions. For ‘every’-SAA, we observed no significant differences with the study of Potts and colleagues (all χ^2’s < 1). In the case of ‘exactly one’-SAA, we observed no significant differences with the study of Potts and colleagues (all p’s $> .10$). In this respect, Exp. 2 diverges from the the P–C–L– basketball task from Exp. 1, which found that the rate of embedded UBCs for ‘exactly one’-SAA was marginally lower than in the study of Potts and colleagues. Nonetheless, there were no significant differences between the rates of embedded UBCs for ‘exactly one’-SAA in Exp. 2 and in the P–C–L– basketball task from Exp. 1 (all p’s $>$ .14). For ‘no’-NAA, the rates of embedded UBCs in Exp. 2 were consistently and significantly lower compared to the results of Potts and colleagues (all p’s $< .05$). This confirms the results that were found for the basketball task from Exp. 1.

In summary, then, the rates of embedded UBCs are relatively unaffected by the presence or absence of a practice phase, the presence or absence of control items, and the length of the experiment. The high rate of embedded UBCs that Potts and colleagues found for ‘no’-NAA (i.e., 43%) was not replicated in any of our tasks, in which the rate of embedded UBCs ranged between 18% and 25%. The high rate of embedded UBCs that Potts and colleagues found for ‘exactly one’-SAA (i.e., 51%), by contrast, was replicated in all tasks other than P–C–L–.

3. Bridging the gap

Taken together, the results of Exps. 1 and 2 largely confirm the pattern of results that was found in the original studies of Geurts and Pouscoulous and Potts and colleagues. That is, in the squares-and-circles task, we observed consistently low rates of embedded UBCs—about 10% in all conditions. In the basketball task, by contrast, the rates of embedded UBCs were substantially higher when ‘some’ was

[3] We initially tried to include item as a random factor. However, this model did not converge.

embedded under ‘exactly one’ (about 40%) or under ‘no’ (about 25%) but not when it was embedded under ‘every’ (about 10%).

These results call for an investigation into the remaining factors that separate the rates of embedded UBC across the two sentence-picture verification tasks. In what follows, we consider four such factors and subject them to an experimental investigation.

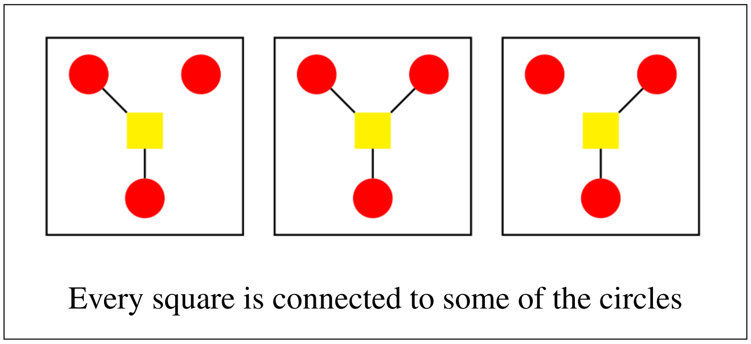
Individuated items The first factor that we consider stems from the following comment from Chemla and Spector (2011, p. 365) about the squares-and-circles task used by Geurts and Pouscoulous:

> [W]e find Geurts and Pouscoulous’ (2009) pictures rather difficult to decipher. Consider the example depicted in [Fig. 2] again. The crucial bit of information for the present purposes is that the square on top of the picture is connected with all the circles, hereby falsifying the local reading. On purely introspective grounds, we find this information pretty hard to extract, and participants may either miss it or ignore it altogether… [W]e might expect that the local reading would be significantly more accessible if we used pictures that prompted subjects to pay attention to the specific properties of each particular square, rather than to more global patterns. We may thus hope that by constructing other sentence-picture matching tasks in which the relevant items are more clearly individuated, we will be able to make the local reading more relevant.

In order to test whether individuated items promote the derivation of embedded UBCs, we conducted a squares-and-circles task using displays with more individuated items, such as the one shown in Fig. 8. If Chemla and Spector’s observation is correct, we should expect a significant increase in the rates of embedded UBCs compared to the original squares-and-circles task. The details of this experiment will be reported in Exp. 3.

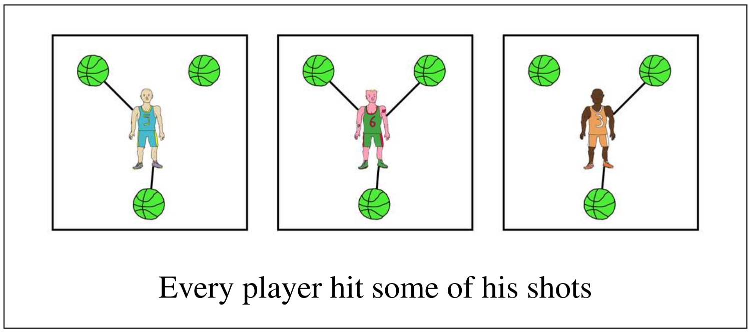
Abstractness The second factor is described in the following remark from Potts and colleagues (2016, p. 779), in which the squares-and-circles task is viewed less favourably than the basketball task:

> First, we should seek out simple, naturalistic stimuli. Previous experiments in this area have used abstract displays. Together with the inevitable complexity of the sentences involved, this choice seems likely to put cognitive



Every square is connected to some of the circles

Figure 8: Example item from Exp. 3 in which the items are clearly individuated.



Every player hit some of his shots

Figure 9: Example item from Exp. 4 which uses displays that are structurally similar but less abstract than the original squares-and-circles task.

> demands on participants in ways that could affect the stability and reliability of the responses.

If this criticism is accurate, we should expect the rates of embedded UBCs to increase if we recreate the squares-and-circles task using players that are connected to basketballs, with connections signifying whether or not the corresponding shot was a hit, as in Fig. 9. After all, such displays are structurally similar but less abstract than the displays used in the squares-and-circles task of Exp. 1. In Exp. 4, we conducted an experiment using such displays.

Subitising The third factor focuses on the role of subitising. Subitising refers to the ability to instantly recognise the numerosity of small groups of objects. Psychonomic research has shown that the subitising range goes from 1 to 4 (e.g., Dehaene, 1997). Degen and Tanenhaus (2011, p. 3300) have argued that there is a connection between subitisability and UBCs:

> Given that subitizing is accompanied by a strong sense of number "popping out", it is conceivable that number terms are more natural labels for set sizes in this range. From a Gricean perspective, speakers should use the most natural label available to establish reference. If instead a vague quantifier like 'some' is used, it is likely that this will result in increased processing effort on the comprehender's part.

While Degen and Tanenhaus are concerned with reference, an analogous case can be made for assertion (cf. Potts et al., 2016, p. 26). Suppose a speaker wants to describe a situation in which, e.g., 3 out of 10 plates are clean. Since she immediately recognises the numerosity of the set of clean plates, she is expected to say 'Three of the plates are clean' rather than the indeterminate 'Some of the plates are clean'. In other words, 'some' is an unnatural description of sets of objects whose numerosity lies within the subitising range.

In the displays used by Geurts and Pouscoulous, 'some' always referred to quantities within the subitising range. Hence, participants might be surprised that the speaker did not use a more precise quantifying expression. To illustrate, in the 'exactly two'-NSSA condition, shown in Fig. 2, the speaker could have used 'two' instead of 'some'. Although no such straightforward alternative is available in the case of 'every'-SSA, the target sentence might still be seen to compete with non-equivalent descriptions such as the following:

(7) a. Two squares are connected with two circles.
b. One square is connected to all circles.

As a consequence of this unnatural use of 'some', participants in the study of Geurts and Pouscoulous might be hesitant to depart from the literal interpretation of the corresponding sentences, since the speaker's behaviour seems at odds with how participants expect a rational speaker to behave. Potts and colleagues, in order to prevent such "competitions from cardinal determiners" (p. 26), hence chose to use displays in which 'some' referred to quantities outside of the subitising range, which renders their results impervious to this criticism.

If this explanation is correct, we should expect that the rate of embedded UBCs in the basketball paradigm will decrease significantly once 'some' refers to quantities that lie within the subitising range. We test this prediction in Exp. 5. In that experiment, we repeated the basketball task while making sure that

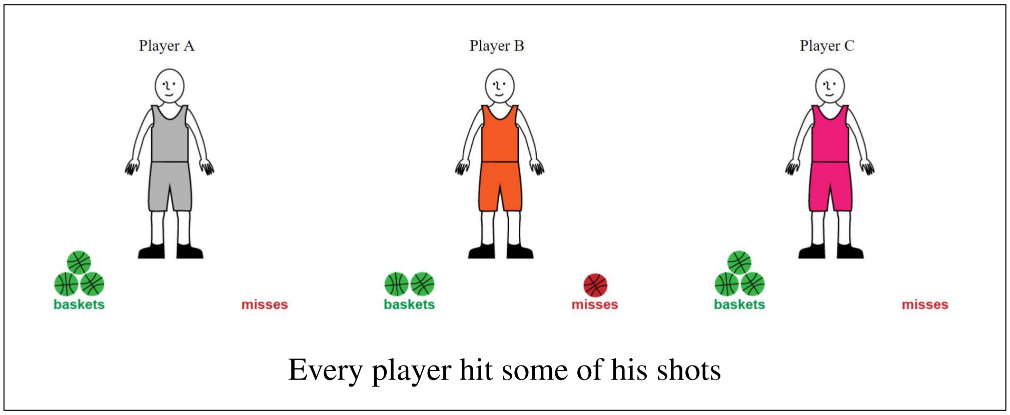


Figure 10: Example item from Exp. 5 in which 'some' refers to quantities that lie within the subitising range (i.e., between 1 and 4).

'some' referred to the same quantities as it did in the squares-and-circles task, using displays such as the one shown in Fig. 10.

All, some, none The fourth factor zooms in on the distribution of A, S, and N cases in the critical items. The squares-and-circles task tested the following critical items: 'every'-SSA, 'exactly one'-NNNA, 'exactly one'-NNSA, and 'no'-NNNA. The basketball task tested the following critical items: 'every'-SAA, 'exactly one'-SAA, and 'no'-NAA. One might hypothesise that the contrast between different types of players or squares (i.e., A, S, and N cases) was more salient in the latter items than in the former.

To illustrate, compare 'exactly one'-SAA and 'exactly one'-NNSA. In the former case, the contrast between the A cases and the S case is intuitively more obvious, since there were only three players who made either some but not all or all of their shots, whereas, in the latter case, there were four players who made either none, some but not all, or all of their shots. The salience of the contrast between different types of players might have increased the contrast between the scalar expressions 'some' and 'all', which, in turn, might have facilitated the derivation of embedded UBCs.

Effects of the distribution of the A, S, and N cases on the frequency of embedded UBCs have previously been reported in the literature. For example, Chemla and Spector (2011) conducted an experiment in which participants had to judge on a graded scale how well sentences described corresponding displays. In one condition, the critical sentence was:

(8) Every letter is connected to some of its circles.

Participants were presented with seven types of displays (where, e.g., 4N2S is shorthand for NNNNSS): 6N, 4N2S, 2N4S, 6A, 2S4A, 4S2A, and 6S. Even though the sentence is unambiguously false in the first three situations, Chemla and Spector found that participants gave lower ratings to 6N than to 4N2S, and lower ratings to 4N2S than to 2N4S. In a similar vein, even though, in both 2S4A and 4S2A, the sentence is true on its literal interpretation or on its interpretation with a global UBC, while it is false if an embedded UBC is derived, participants gave lower ratings to 2S4A than to 4S2A. Van Tiel (2014) explains these differences in terms of typicality: 6S is the prototype associated with (8) and the ratings that participants give reflect the distance from this prototype.

Geurts and van Tiel (2013) conducted a similar experiment testing how well the following sentence described various displays:

(9) Exactly one letter is connected to some of its circles.

Participants in their experiment provided significantly lower ratings for NSA than for SAA, despite the fact that, in both situations, the sentence is false on its literal interpretation and on its interpretation with a global UBC, but true if the embedded UBC is derived. Geurts and van Tiel argue that the contrast between S cases and A cases is substantially more salient in the SAA situation than in the NSA situation, since the latter situation contains three types of letters. This, in turn, might have made the contrast between ‘some’ and ‘all’ more prominent in the SAA situation, thus causing participants to interpret ‘some’ as excluding ‘all’.

Note, however, that Potts and colleagues tested both ‘exactly one’-NSA and ‘exactly one’-SAA, but failed to find a difference in the proportion of embedded UBC responses. Indeed, numerically, the rate of embedded UBC responses was higher in the former case than in the latter.

Nonetheless, the results of Potts and colleagues also provide indications that the distribution of A, S, and N cases affects the rates of embedded UBCs. For example, ‘no’-NAA led to significantly higher rates of embedded UBCs than ‘no’-NNA, which in turn prompted significantly higher rates of embedded UBCs than ‘no’-AAA. In all of these cases, the sentence is literally false but true if interpreted with an embedded UBC, suggesting that the rates of embedded UBC responses might be influenced by the precise distribution of A, S, and N cases.

If this explanation is on the right track, we should expect that the rate of embedded UBCs in the basketball task will decrease significantly if the same distributions of A, S, and N cases are used as in the squares-and-circles task. For

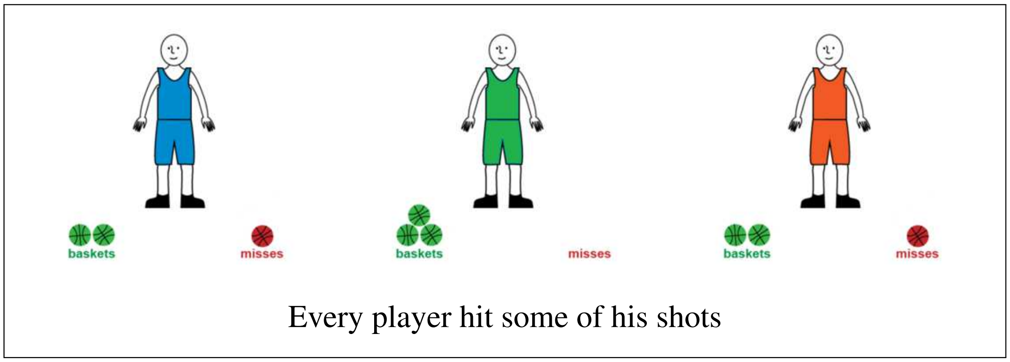


Figure 11: Example item from Exp. 6 in which the same critical items are tested as in the original squares-and-circles task.

example, Fig. 8 shows an 'every'-SSA item instead of the 'every'-SAA item tested in the original basketball task. In Exp. 6, we test whether changing the distribution of A, S, and N cases affects the rates of embedded UBCs.

Fig. 12 provides an overview of the hypotheses that will be tested in Exps. 3 to 6. In the following sections, we describe these experiments in more detail.

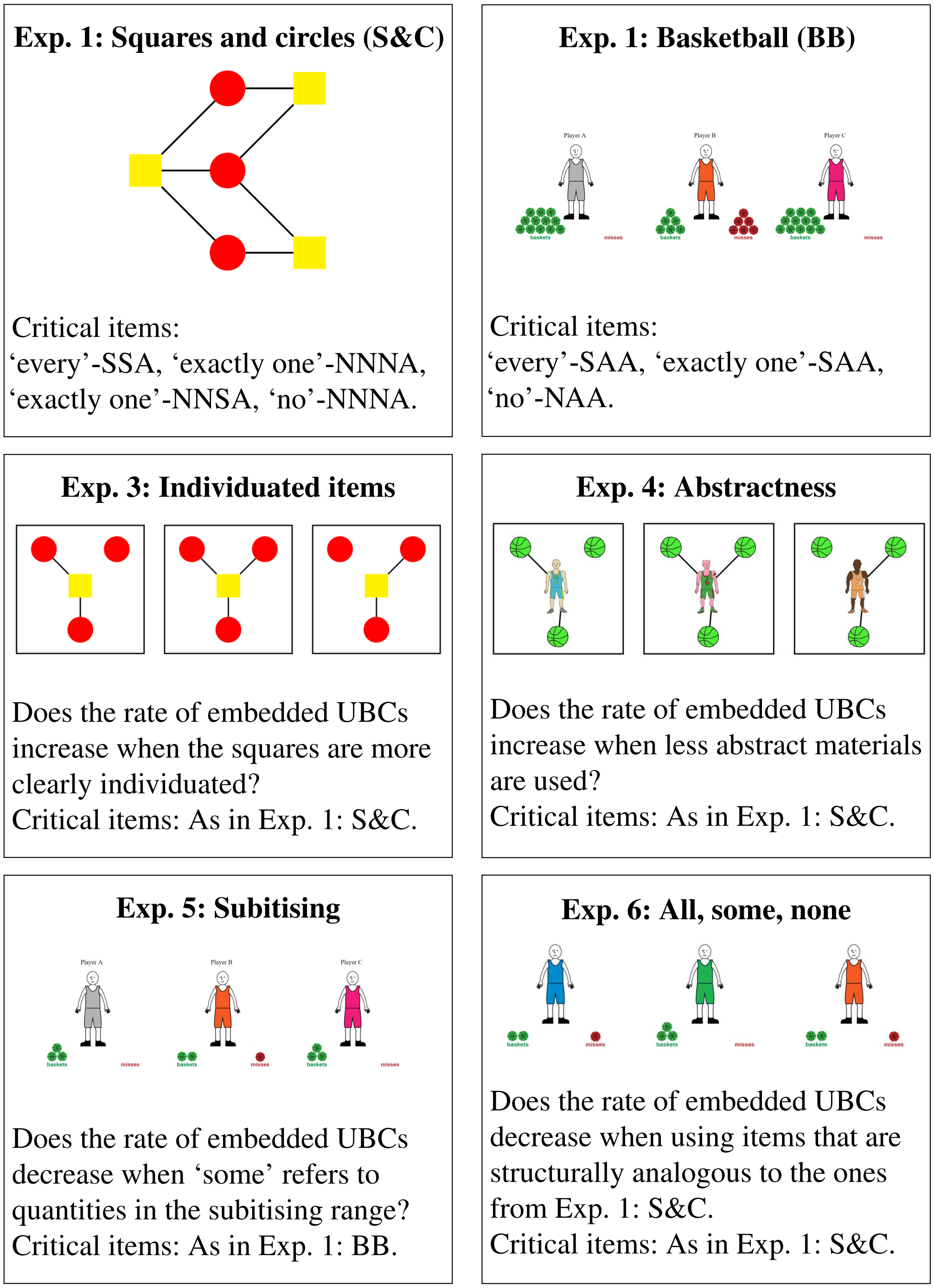
Experiment 3: Individuated items

Participants

120 participants were drafted on Mechanical Turk. 7 participants were removed from the analysis either because their native language was not English or because they had participated in a similar experiment before. The mean age of the remaining participants was 35 (standard deviation: 11). 52 of the participants were female.

Materials

The materials were identical to the ones used in the squares-and-circles task reported in Exp. 1. That is, the experiment consisted of 4 target items ('every'-SSA, 'exactly one'-NNSA, 'exactly one'-NNNA, and 'no'-NNNA), interspersed with 13 fillers. The only difference with the original squares-and-circles task concerned the structure of the displays. In the original squares-and-circles task, displays showed squares and circles on a partially filled 3×3 grid; in this experiment, the squares were shown in separate boxes. See Fig. 8 for an example item.



Exp. 1: Squares and circles (S&C)

Critical items:
'every'-SSA, 'exactly one'-NNNA, 'exactly one'-NNSA, 'no'-NNNA.

Exp. 1: Basketball (BB)

Critical items:
'every'-SAA, 'exactly one'-SAA, 'no'-NAA.

Exp. 3: Individuated items

Does the rate of embedded UBCs increase when the squares are more clearly individuated?
Critical items: As in Exp. 1: S&C.

Exp. 4: Abstractness

Does the rate of embedded UBCs increase when less abstract materials are used?
Critical items: As in Exp. 1: S&C.

Exp. 5: Subitising

Does the rate of embedded UBCs decrease when 'some' refers to quantities in the subitising range?
Critical items: As in Exp. 1: BB.

Exp. 6: All, some, none

Does the rate of embedded UBCs decrease when using items that are structurally analogous to the ones from Exp. 1: S&C.
Critical items: As in Exp. 1: S&C.

Figure 12: Overview of the displays and hypotheses to be tested in Exps. 3–6.

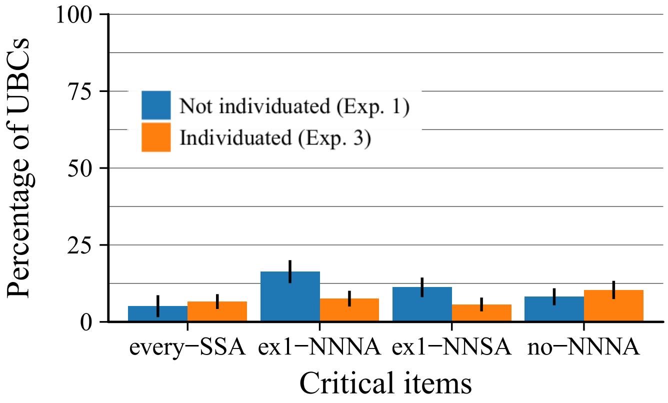


Figure 13: Percentage of embedded UBC responses in the squares-and-circles task depending on whether the items were individuated (Exp. 3) or not (Exp. 1). Error bars represent standard errors.

Procedure

The procedure was the same as in the original squares-and-circles task.

Results

6 participants were removed from the analysis for making mistakes in more than 20% of the filler items. The mean error rate of the remaining participants was 3.2%. The percentages of UBC responses for each critical item are shown in Fig. 13 alongside the previously reported results for the original squares-and-circles task.

The average rate of UBC responses in Exp. 3 (7.5%) was numerically lower than the average rates of UBC responses in the original squares-and-circles (10.2%) and basketball (25.3%) tasks reported in Exp. 1. To determine whether these differences were significant, we constructed a mixed effects logistic regression model predicting response (UBC or other) on the basis of task (Exp. 1: Squares-and-circles, Exp. 1: Basketball, or Exp. 3), with random intercepts for participants and items. Exp. 3 was used as reference level for the task factor. The rate of embedded UBCs in Exp. 3 did not differ from the original squares-and-circles task (β = 0.36, SE = 0.32, Z = 1.15, p = .25) but was significantly lower than in the original basketball task (β = -1.68, SE = 0.32, Z = -5.26, p < .001).

Hence, the results of Exp. 3 show that the frequency of embedded UBCs is independent of whether or not the items are clearly individuated.

Experiment 4: Abstractness

Participants

120 participants were drafted on Mechanical Turk. 13 participants were removed from the analysis either because their native language was not English or because they had participated in a similar experiment before. The mean age of the remaining participants was 36 (standard deviation: 10). 48 of the participants were female.

Materials

The critical items were those used in the original squares-and-circles task reported in Exp. 1: ‘every’-SSA, ‘exactly one’-NNSA, ‘exactly one’-NNNA, and ‘no’-NNNA. Critical items were interspersed with 13 fillers, which were the same as those used in the original basketball task. The displays were structurally analogous to the ones used in Exp. 3. Rather than using squares and circles, however, the displays consisted of interconnected basketball players and basketballs. The number of players, basketballs, and connections was the same as the number of squares, circles, and connections in Exp. 3.

Procedure

The procedure was the same as in the original basketball task reported in Exp. 1. That is, the instructions indicated that participants were to see computer-generated commentary on basketball free-throw shooting competitions. The only difference was that, in the original basketball task, hits and misses were represented by two piles of basketballs next to the players. In this experiment, hits were represented by connections and misses by the absence of connections.

Results

11 participants were removed from the analysis for making mistakes in more than 20% of the filler items. The mean error rate of the remaining participants was 4.7%. The percentages of UBC responses for each critical item are shown in Fig. 14 alongside the previously reported results of the original squares-and-circles task.

The mean rate of UBC responses in Exp. 4 (10.9%) was almost the same as in the original squares-and-circles task (10.2%) and lower than in the original basketball task (25.3%). To determine whether these differences were significant, we constructed a mixed effects logistic regression model predicting response (UBC

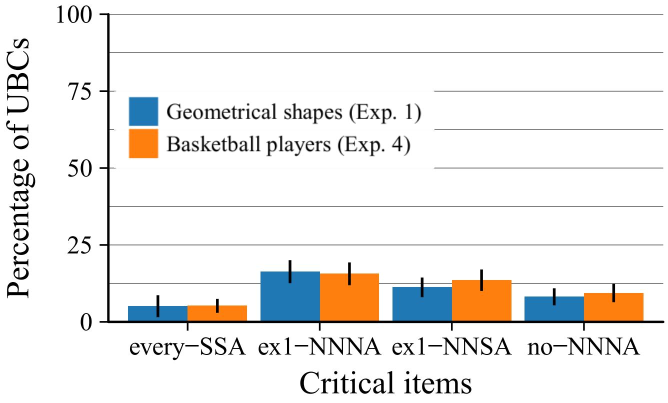


Figure 14: Percentage of embedded UBC responses depending on whether the displays showed geometrical shapes (Exp. 1) or basketball players (Exp. 4). Error bars represent standard errors.

or other) on the basis of task (Exp. 1: Squares-and-circles, Exp. 1: Basketball, or Exp. 4), with random intercepts for participants and items. Exp. 4 was used as reference level for the task factor. The rate of embedded UBCs in Exp. 4 did not differ from the original squares-and-circles task (β = -0.04, SE = 0.34, $Z >$ -1) but was significantly lower than in the original basketball task (β = 1.50, SE = 0.35, Z = 4.30, $p < .001$).

In summary, the results of Exp. 4 show that the frequency of embedded UBCs is independent of whether the experiment is about connected geometrical shapes or about basketball players in a free-throw shooting competition.

These results should not be construed as conclusively showing that the prevalence of embedded UBCs is independent of the abstractness of the displays: there might be other manipulations of abstractness that do affect the frequency of embedded UBCs. For example, participants might be more likely to derive embedded UBCs when hits and misses are distinguished on the basis of the colour of the balls rather than the presence or absence of connections between players and balls. However, our particular manipulation of abstractness clearly did not affect the frequency of embedded UBCs.

The hypothesis that abstractness influences the rate of embedded UBCs, as formulated by Potts and colleagues, rests on two assumptions: (i) the derivation of embedded UBCs is cognitively costly and (ii) the prevalence of embedded UBCs therefore decreases with the difficulty of the experimental task. In order to evaluate

Experiment	RT (F)	RT (T)	UBC
Exp. 1: Squares	7167	6418	9.8
Exp. 1: Basketball	7012	6768	24.8
Exp. 2	5864	5749	24.4
Exp. 3: Individuated	5403	5710	7.5
Exp. 4: Abstract	5268	4816	11.1
Exp. 5: Subitising	5935	5961	20.5
Exp. 6: All, some, none	5235	5659	16.5

Table 3: Mean response times in milliseconds for filler (F) and target (T) items and percentages of embedded UBC responses in each experiment.

these two assumptions, we group together all of the data that are presented in this paper, removing responses that took longer than 20 seconds (2.9% of the data).

In line with the assumption that the derivation of embedded UBCs is cognitively costly, embedded UBC responses (6542 milliseconds on average) took longer than other responses (5655 milliseconds on average). In order to determine if this difference was statistically significant, we constructed a mixed effects linear regression model predicting logarithmised response times based on response (UBC or other) with random intercepts for participants, items, and experiments. The model confirmed that UBC responses were significantly slower than other responses ($\beta = 0.08$, $SE = 0.04$, $t = 2.19$, $p = .03$).

It would be a mistake, however, to conclude from this observation that the prevalence of embedded UBCs is a function of the difficulty of the task, as Potts and colleagues assume. To illustrate, Table 3 shows the mean response times in filler and target items and the corresponding rates of embedded UBC responses for each experiment. If anything, higher response times to filler ($r(5) = .25$, $t < 1$) and target ($r(5) = .39$, $t < 1$) items were associated with higher rates of embedded UBC responses, although neither correlation was statistically significant.

In a similar vein, participants who found the task more difficult, as measured in terms of their response times to filler items, were not less likely to derive embedded UBCs. To show this, we determined for each participant their average response time to filler items, as well as the number of embedded UBC responses. There was no significant correlation between these two measures ($r(2050) = .03$, $t = 1.30$, $p = .19$). That is, participants who found the task more difficult were neither more nor less likely to derive embedded UBCs.

Finally, embedded UBC responses were not more prevalent in less difficult conditions, as measured in terms of response times. Indeed, embedded UBCs were more prevalent for ‘exactly one’-SAA/NNSA (23%) and ‘no’-NAA/NNNA (16%) than for ‘every’-SAA/SSA (8%) and ‘exactly one’-NNNA (14%), even though the former conditions were associated with higher response times. To show this, we constructed a mixed effects linear regression model predicting logarithmised response times based on condition (‘every’-SAA/SSA, ‘exactly one’-SAA/NNSA, ‘exactly one’-NNNA, or ‘no’-NAA/NNNA), with random intercepts for participants, experiments, and responses. Multiple comparisons were carried out based on Tukey’s procedure, as implemented in the `multcomp` package (Hothorn, Bretz, & Westfall, 2008). These comparisons indicated that response times were significantly faster for ‘every’-SAA/SSA than for all other conditions (all *p*’s $< .001$) and significantly faster for ‘exactly one’-NNNA than for ‘exactly one’-SAA/NNSA and ‘no’-NAA/NNNA. This response time pattern speaks against the idea that embedded UBCs are more prevalent in less difficult conditions.

Taken together, these observations indicate that the frequency of embedded UBCs is independent of the difficulty of the experimental task.

Experiment 5: Subitising

Participants

120 participants were drafted on Mechanical Turk. 25 participants were removed from the analysis either because their native language was not English or because they had participated in a similar experiment before. The mean age of the remaining participants was 32 (standard deviation: 9). 33 of the participants were female.

Materials

The materials were identical to the basketball task reported in Exp. 1. That is, the experiment consisted of 3 critical items (‘every’-SAA, ‘exactly one’-SAA, and ‘no’-NAA), interspersed with 13 fillers, which were identical to the ones used in the original basketball task. The only difference with the original basketball task was the number of free throws players attempted. In the original task, players always attempted 12 free throws. In this experiment, the number of free throws players attempted was matched with the number of circles in the displays used in the squares-and-circles task. This number ranged from 2 to 4 and was thus always in the subitising range. An example item is shown in Fig. 10.

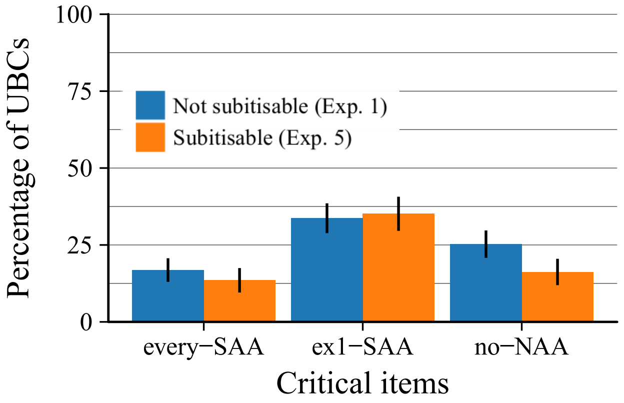


Figure 15: Percentage of embedded UBC responses in the basketball task depending on whether ‘some’ referred to quantities inside (Exp. 5) or outside (Exp. 1) the subitising range. Error bars represent standard errors.

Procedure

The procedure was the same as in the original basketball task reported in Exp. 1.

Results

17 participants were removed from the analysis for making mistakes in more than 20% of the filler items. The mean error rate of the remaining participants was 3.3%. The percentages of UBC responses for each critical item are shown in Fig. 15 alongside the results of the original basketball task.

The mean rate of UBC responses in Exp. 5 (21.6%) was almost the same as in the original basketball task (25.3%) and higher than in the original squares-and-circles task (10.2%). To determine whether these differences were significant, we constructed a mixed effects logistic regression model predicting response (UBC or other) on the basis of task (Exp. 1: Squares-and-circles, Exp. 1: Basketball, or Exp. 5), with random intercepts for participants and items. Exp. 5 was used as reference level for the task factor. The rate of embedded UBCs in Exp. 5 did not differ from the original basketball task (β = 0.26, SE = 0.32, $Z < 1$) but was significantly higher than in the original squares-and-circles task (β = -1.26, SE = 0.36, Z = -3.48, $p < .001$).

In summary, the results of Exp. 5 show that the frequency of embedded UBCs is independent of whether or not ‘some’ refers to quantities in the subitising range.

Experiment 6: All, some, none

Participants

120 participants were drafted on Mechanical Turk. 27 participants were removed from the analysis either because their native language was not English or because they had participated in a similar experiment before. The mean age of the remaining participants was 36 (standard deviation: 11). 44 of the participants were female.

Materials

The materials were similar to the original basketball task. However, Exp. 6 tested the 4 critical items that were tested in the original squares-and-circles task (i.e., 'every'-SSA, 'exactly one'-NNSA, 'exactly one'-NNNA, and 'no'-NNNA) instead of the 3 critical items that were tested in the original basketball task (i.e., 'every'-SAA, 'exactly one'-SAA, and 'no'-NAA). In addition, the number of free throws that players attempted was matched with the number of circles in the displays used in the squares-and-circles task. As a consequence, the number of free throws that players attempted was always in the subitising range. An example item is shown in Fig. 11.

Procedure

The procedure was the same as in the original basketball task.

Results

4 participants were removed from the analysis for making mistakes in more than 20% of the filler items. The error rate of the remaining participants was 3.6%. The percentages of UBC responses for each critical item are shown under 'Original study' in Fig. 16 alongside the results of the original squares-and-circles and basketball tasks from Exp. 1.

To analyse the results, we constructed a mixed effects logistic regression model predicting response (UBC or other) on the basis of task (Exp. 1: Squares-and-circles, Exp. 1: Basketball, or Exp. 6) with random intercepts for participants and items. Exp. 6 was used as reference level for the task factor. We observed that the percentage of embedded UBC responses in Exp. 6 (16.6%) was significantly lower than in the original basketball task (25.3%) (β = -0.79, *SE* = 0.32, *Z* = -2.51, *p* =

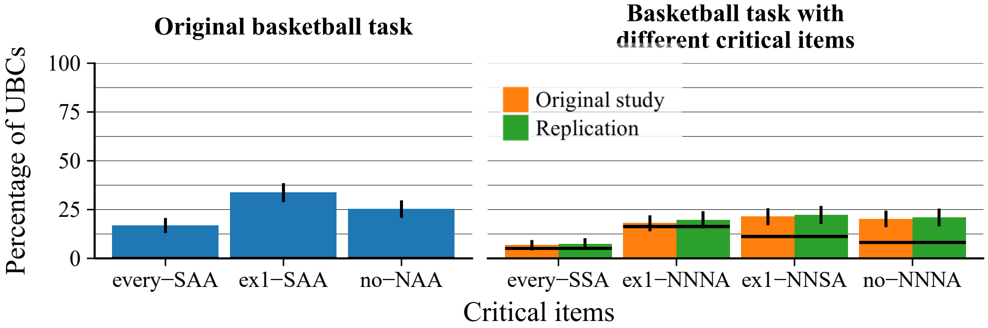


Figure 16: Percentage of embedded UBC responses in the basketball task using the critical items from the squares-and-circles task (Exp. 6 and its replication), alongside the percentages of UBC responses in the original basketball task (Exp. 1). Horizontal lines represent the percentages of embedded UBC responses in the original squares-and-circles task (Exp. 1). Error bars represent standard errors.

.03) and was not significantly different from the original squares-and-circles task (10.2%) ($\beta = 0.67$, $SE = 0.32$, $Z = 2.07$, $p = .10$).

Focusing on the comparison with the original basketball task, chi-square tests for proportions indicate that the percentage of embedded UBC responses for ‘every’-SSA was significantly lower in Exp. 6 (6.7%) than for ‘every’-SAA in the original basketball task (17.1%, $\chi^2 = 4.81$, $p = .03$). The percentage of embedded UBC responses for ‘exactly one’-NNSA was significantly lower in Exp. 6 (21.3%) than for ‘exactly one’-SAA in the original basketball task (36.2%, $\chi^2 = 5.11$, $p = .02$). The percentage of embedded UBC responses for ‘no’-NNNA in Exp. 6 (20.2%) did not differ significantly from the percentage of embedded UBC responses for ‘no’-NAA in the original basketball task (24.8%, $\chi^2 < 1$).

The observation that the rate of embedded UBC responses for ‘every’-SSA in Exp. 6 was significantly lower than for ‘every’-SAA in the original basketball task is in line with the aforementioned observation that participants in Chemla and Spector’s (2011) experiment gave lower ratings to ‘Every letter is connected to some of its circles’ in the 4A2S situation than in the 2A4S situation. In both studies, the rates of responses that suggest that participants derived an embedded UBC are higher for situations that are further from the prototypical situation, i.e., SSS in our experiments and 6S in Chemla and Spector’s study.

Focusing on the comparison with the original squares-and-circles task, chi-square tests for proportions indicate that the percentage of embedded UBC responses for ‘every’-SSA and ‘exactly one’-NNNA were not significantly different across the two tasks (6.7% and 18.0% in Exp. 6 vs. 5.1% and 16.3% in the original squares-and-circles task, both χ^2’s < 1). The percentage of embedded UBC responses for ‘exactly one’-NNSA was marginally lower in the original squares-and-circles task (11.2%) than in Exp. 6 (21.4%, $\chi^2 = 3.5$, $p = .06$). The percentage of UBC responses for ‘no’-NNNA was significantly lower in the original squares-and-circles task (8.2%) than in Exp. 6 (20.2%, $\chi^2 = 5.64$, $p = .02$).

One might be concerned that, giving the large number of experiments, one of the analyses was bound to come out significant because of chance rather than the presence of an actual effect in the population. In order to alleviate this concern, we conducted an exact replication of Exp. 6.

120 participants (mean age: 33; standard deviation: 10; 54 females) were drafted on Mechanical Turk. 26 participants were removed from the analysis for having participated in a similar experiment before or for not having English as their native language. A further 13 participants were removed from the analysis for making errors in more than 20% of the filler items. The percentage of embedded UBC responses (16.9%) in the replication was almost identical to the original experiment (16.6%). The percentages of UBC responses for each critical item are shown under “Replication” in Fig. 16 alongside the results of the original squares-and-circles and basketball tasks from Exp. 1.

To analyse the results, we constructed a mixed effects logistic regression model predicting response (UBC or other) on the basis of experiment (Exp. 1: Squares-and-circles, Exp. 1: Basketball, or replication of Exp. 6) with random intercepts for participants and items. The replication of Exp. 6 was used as reference level for the task factor. In line with the original study, we observed that the percentage of embedded UBC responses in the replication of Exp. 6 (16.6%) was significantly lower than in the original basketball task (25.3%) ($\beta = 0.70$, $SE = 0.32$, $Z = 2.22$, $p = .03$). Unlike the original study of Exp. 6, the percentage of embedded UBC responses in the replication of Exp. 6 was significantly higher than in the original squares-and-circles task (10.2%) ($\beta = -0.76$, $SE = 0.32$, $Z = -2.34$, $p = .02$).

Focusing on the comparison with the original basketball task, chi-square tests for proportions indicate that the percentage of embedded UBC responses for ‘every’-SSA was marginally lower in the replication of Exp. 6 (7.4%) than for ‘every’-SAA in the original basketball task (17.1%, $\chi^2 = 3.71$, $p = .05$). The percentage of embedded UBC responses for ‘exactly one’-NNSA was significantly lower in the replication of Exp. 6 (22.2%) than for ‘exactly one’-SAA in the original basketball task (36.2%, $\chi^2 = 4.08$, $p = .04$). The percentage of embedded UBC responses for ‘no’-NNNA in the replication of Exp. 6 (20.9%) did not differ significantly from the percentage of embedded UBC responses for ‘no’-NAA in the original basketball task (24.8%, $\chi^2 < 1$).

Focusing on the comparison with the original squares-and-circles task, chi-square tests for proportions indicate that the percentage of embedded UBC responses for ‘every’-SSA’, ‘exactly one’-NNNA, and ‘exactly one’-NNSA were not significantly different across the two tasks (7.4%, 22.2%, and 19.7% in the replication of Exp. 6 vs. 5.1%, 16.3%, and 11.2% in the original squares-and-circles task, all p’s $> .11$). The percentage of UBC responses for ‘no’-NNNA was significantly lower in the original squares-and-circles task (8.2%) than in the replication of Exp. 6 (20.9%, $\chi^2 = 5.93$, $p = .01$).

In summary, then, using displays with different distributions of A, S, and N cases has a significant effect on the frequency of embedded UBCs. At the same time, this factor does not explain all of the differences between the original basketball and squares-and-circles tasks, in particular when it comes to the ‘exactly one’-NNSA and ‘no’-NNNA conditions.

4. General discussion

4.1. Summary

We started this paper with a brief discussion of the Wason Selection Task, in which participants have to evaluate conditional rules such as ‘If there is a vowel on one side, then there is an even number on the other’. The selection task has engendered a prodigous amount of debate. The main lesson to be learned from this debate, from our perspective, is that task performance is largely shaped by features of the materials that are used.

The debate about the Wason Selection Task shows some marked similarities with the debate about ‘embedded implicatures’, which revolves around the question of whether ‘some’ is interpreted as ‘some but not all’ when it is embedded under a quantifying expression. In their pioneering study, Geurts and Pouscoulous (2009) observed negligible rates of such embedded UBCs. In a follow-up study that modifies the experimental paradigm, however, Potts and colleagues (2016), observed more substantial rates of embedded UBCs in at least some embedding environments.

The two studies differ in a large number of methodological respects, which complicates a straightforward comparison. In order to level the playing field, Exp. 1 repeated both studies in a more uniform experimental setting. For convenience, we refer to these quasi-replications as the *squares-and-circles task* (Geurts and Pouscoulous) and the *basketball task* (Potts and colleagues). The results confirmed that participants were significantly less likely to derive embedded UBCs in the squares-and-circles task than in the basketball task. The difference, however, was less pronounced than in the original studies.

In particular, the rate of embedded UBCs in the basketball task was significantly lower than in the original study of Potts and colleagues. Exp. 2 explores three differences between the basketball task and the original study of Potts and colleagues that might underlie this discrepancy. First, the original study but not the basketball task was preceded by a practice phase. Second, the original study also tested sentences with 'all' and 'none' instead of 'some'. Third, the original study was twice as long as the basketball task. We observed that these three factors are responsible for the different rates of embedded UBCs for 'exactly one'-SAA but that the rate of embedded UBCs for 'no'-NAA was consistently lower than in the original study of Potts and colleagues.

Afterwards, we evaluated four possible explanations for the different rates of embedded UBCs in the squares-and-circles and basketball tasks. The first two of these candidate explanations were derived from criticisms of the study of Geurts and Pouscoulous that had been voiced earlier in the literature.

According to the first of these criticisms, the displays used in the study of Geurts and Pouscoulous were too difficult to parse because the shapes were not clearly individuated. In order to evaluate this criticism, Exp. 3 repeated the squares-and-circles task using displays with shapes that were more clearly individuated. This manipulation, however, did not increase the rate of embedded UBCs.

According to the second criticism, the displays used in the study of Geurts and Pouscoulous were too abstract. The processing of these displays thus demanded too many cognitive resources for participants to be able to derive embedded UBCs. In order to evaluate this criticism, Exp. 4 repeated the squares-and-circles task using displays with basketball players instead of geometrical shapes. Again, however, this manipulation did not increase the rate of embedded UBCs. Indeed, although an analysis of the reaction times indicated that the derivation of embedded UBCs was associated with a processing cost, there was no straightforward relationship between the difficulty of the experimental task and the rate of embedded UBCs.

Taken together, the results of Exps. 3 and 4 caution against giving too much credence to armchair criticisms of experimental data, and suggest that it is more fruitful to tease apart the relevant methodological features that are responsible for discrepant results—as researchers on the Wason Selection Task did over the years. Hence, in Exps. 5 and 6, we considered the impact of two methodological factors.

Exp. 5 tested if it makes a difference whether ‘some’ refers to quantities within or outside of the subitising range (i.e., between 1 and 4). Previous research has shown that people find ‘some’ unnatural as a description of quantities in the subitising range (e.g., Degen & Tanenhaus, 2011; van Tiel, 2014). Since, in the squares-and-circles task, ‘some’ always referred to quantities in the subitising range, participants in that task might have been too distracted to derive embedded UBCs. To test this explanation, Exp. 5 repeated the basketball task, making sure that ‘some’ always referred to quantities within the subitising range. This manipulation, however, did not affect the rate of embedded UBCs.

Finally, Exp. 6 investigated whether the precise configuration of the displays influences the rates of embedded UBCs. In both the squares-and-circles task and the basketball task, the displays consisted of N-cases (squares connected to none of the circles or basketball players who made none of their shots), S-cases (squares connected to some but not all of the circles or basketball players who made some but not all of their shots), and A-cases (squares connected to all of the circles or basketball players who made all of their shots). The precise distribution of these cases, however, differed in both tasks. For example the squares-and-circles task tested ‘exactly one’-NNSA, whereas the basketball task tested ‘exactly one’-SSA. To determine whether this difference influences the rate of embedded UBCs, Exp. 6 repeated the basketball task, using the same display configurations as in the squares-and-circles task. This manipulation significantly decreased the rate of embedded UBCs.

However, the rate of embedded UBCs in Exp. 6 was in some conditions still higher than in the squares-and-circles task. So there must be other factors that influence the rate of embedded UBCs. Alternatively, it might be the case that the factors that we manipulated in Exps. 3–5 do in fact have a significant effect, but that the effect was too small to be detected with the sample sizes that we tested. However, we must leave this issue to further research.

Fig. 17 provides a visual overview of the overall rates of embedded UBCs in each condition and experiment.

4.2. Theoretical consequences

In the introduction, we distinguished two routes for scalar expressions to arrive at an upper-bounded reading. First, UBCs can be the result of Gricean reasoning about the speaker’s beliefs and intentions. Second, UBCs can be due to truth-conditional narrowing, which involves restricting the semantic interpretation of the scalar expression.

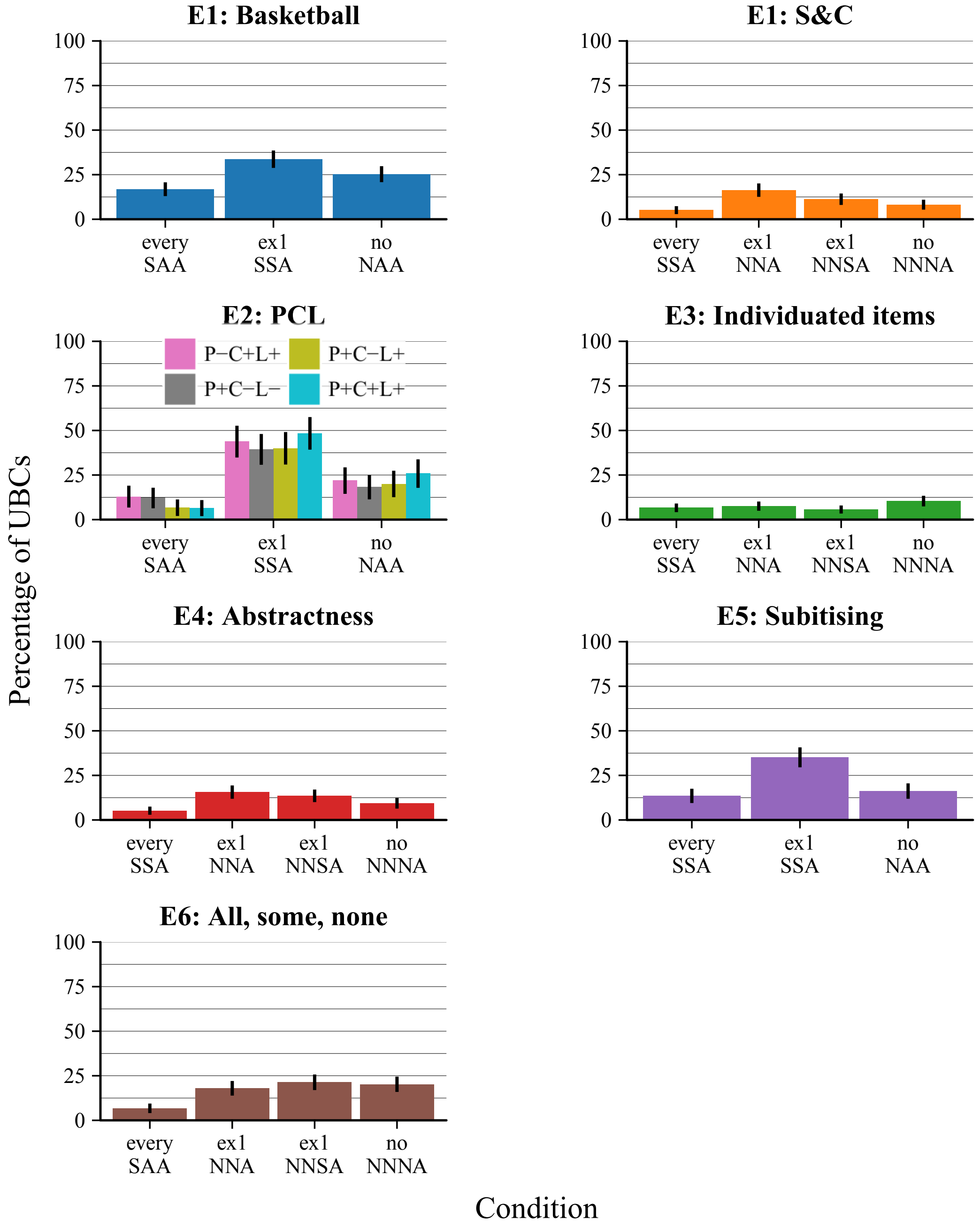


Figure 17: Percentage of UBC responses for each experiment and condition. Error bars represent standard errors.

While there is general agreement that both pragmatic and truth-conditional mechanisms variously underlie UBCs, there has been substantial debate about the division of labour between them. Pragmatic theories, on the one hand, assume that Gricean reasoning is the norm and that truth-conditional narrowing is a fringe phenomenon requiring prosodic marking or a prominent contrast with other scalar expressions. Conventionalist theories, on the other hand, argue that truth-conditional narrowing is the norm and that the scope of Gricean reasoning is limited.

The two types of theories make divergent predictions about how frequently scalar expressions should receive an upper-bounded reading when they occur in the scope of a quantifying expression. Pragmatic theories predict that such embedded UBCs should be rare, while conventionalist theories predict that they should be common.

Our results pose a challenge to both approaches. On the one hand, pragmatic theories have to account for the substantial rates of embedded UBCs when 'some' is embedded under 'exactly one' (about 40% in our basketball task) and, to a lesser extent, when 'some' is embedded under 'no' (about 25% in our basketball task).

To some extent, these embedded UBC responses might be simple errors. Indeed, across all experiments, we observed a significant negative correlation between performance on filler items and embedded UBC responses ($r(650) = -.15$, $p < .001$). In other words, participants who made more errors on filler items also provided more embedded UBC responses. This correlation suggests that there is nothing purposeful about a small portion of the embedded UBC responses.

Moreover, as Potts and colleagues acknowledge, 'some' is a positive polarity item and its occurrence under a negative element is marked. Indeed, 'some' can be used in negative environments to signal a contrast with its scalemate 'all', as the following sentence illustrates:

(10) He didn't pass SOME of his exams. He passed ALL of them.

Hence, the use of 'some' embedded under 'no' might have evoked the contrast with 'all'. All accounts agree that truth-conditional narrowing, and hence an embedded UBC, is expected under such circumstances. In that respect, it will be interesting to see if the results for 'no' are replicated when using an existential quantifier that is not a positive polarity item, such as 'enkele' in Dutch or 'quelques' in French.

The most stubborn case clearly involves 'some' under 'exactly one'. One mitigating factor for that particular case is that it has been shown that 'exactly one' is sometimes interpreted as 'at least one'. To illustrate, Franke, Schlotterbeck, and

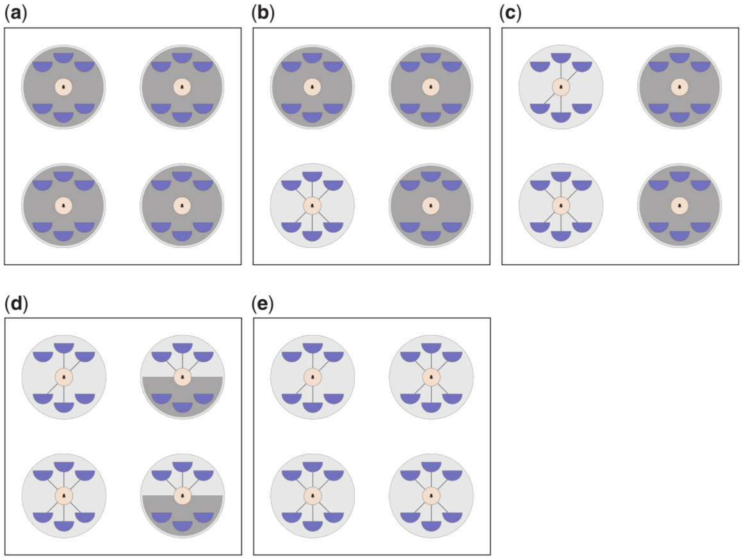


Figure 18: Displays used in the covered picture task reported in Franke et al. (2017). The dark grey area is covered; the light grey area uncovered.

Augurzky (2017) conducted a covered picture task in which participants heard the following sentence along with display (a) shown in Fig. 18:

(11) Exactly one bell is connected with some of its semicircles.

Participants had to indicate whether they were able to decide on the truth value of the sentence. If so, the trial was over. If not, more of the display was uncovered, as shown in display (b), and participants were asked again whether they were able to decide on the truth value of the sentence. If so, the trial was over. If not, more of the display was uncovered, as shown in display (c), and so on. Once the entire display was uncovered, as shown in display (e), participants were forced to indicate whether the sentence was true or false.

Franke and colleagues found that, in about 9% of the trials, participants judged the sentence true at display (b), despite the fact that the truth value of the sentence is still uncertain at this point. After all, there might be another bell that is connected to some of its semicircles. It seems that these participants interpreted ‘exactly one’ as ‘at least one’. Since display (b) reveals that there is one bell connected to some of its semicircles, according to these participants, the sentence can thus be judged true (cf. also Tian et al. 2012).

In a similar vein, participants in our experiments might have interpreted ‘exactly one’ as ‘at least one’, which would account for a portion of the embedded UBCs in this condition. Further research, however, is needed to determine if, and if so, under what circumstances ‘at least’ interpretations of ‘exactly’ occur in sentence-picture verification tasks. To this end, one might test sentences with unambiguous predicates such as ‘Exactly one square is red’ in displays with more than one red square.

Another factor leading to high rates of embedded UBC responses for ‘exactly one’ might be that non-monotonic quantifying expressions such as ‘exactly one’ are inherently contrastive. That is, when evaluating a statement of the form ‘Exactly one player hit some of his shots’, participants need to find one player who hit some of his shots, as well as ascertaining that all other players did not hit some of their shots. Hence, ‘exactly one’ renders salient the contrast between ‘some’ and other scalar expressions, such as ‘all’, and calls upon participants to detect features in which the items in the display differ from each other. This might have caused participants to reinterpret ‘some’ as excluding ‘all’.

This line of argument also explains why the effect of changing the distribution of A, S, and N cases had the greatest effect in the ‘exactly one’ condition: since this was the condition in which participants were concentrated on detecting differences, it was also the condition that was the most sensitive to variations in the salience of these differences. In the other conditions, participants as a rule were not sensitive to differences between items and, hence, varying the salience in those conditions did not have as great an effect. If this line of argument is on the right track, other manipulations that influence the contrast between S-cases and A-cases—e.g., separating the cases based on colour or distance—should have similar effects on the prevalence of embedded UBCs.

In summary, the challenge for pragmatic accounts is to explain the substantial rates of embedded UBCs when ‘some’ is embedded under ‘exactly one’ or ‘no’. We have proposed a number of possible explanations. However, we must leave to further experimental research the questions (i) whether these explanations are on the right track, and (ii) if so, whether they account for a sufficiently large portion of the observed rates of embedded UBCs to relegate the remaining embedded UBCs to the marked operation of truth-conditional narrowing.

Our results are also problematic for conventionalist theories. As discussed in the introduction, one of the challenges for conventionalist theories is to describe the circumstances under which the meaning of ‘some’ is narrowed to ‘some but not all’. The most straightforward proposal holds that ‘some’ is interpreted as ‘some but not all’ across the board (e.g., Ariel, 2015, Levinson, 2000). Other proposals hold that ‘some’ is interpreted with an upper bound if this leads to a logically stronger interpretation (Chierchia, 2006, Chierchia et al., 2012) or to an interpretation that is not logically weaker (Chemla & Spector, 2011).

What all of these positions agree on, then, is that ‘some’ should normally be interpreted as ‘some but not all’ if it is embedded under ‘every’, since, in that case, the embedded UBC leads to a logically stronger interpretation. Across all experiments, this prediction was shown to be mistaken: ‘some’ under ‘every’ was consistently the least likely to be interpreted with an embedded UBC.

Based on this finding, one might consider adopting a constraint according to which ‘some’ is interpreted as ‘some but not all’ if it leads to an interpretation that is not logically stronger. Such a constraint might be partly grounded in a more general preference for truth (Meyer & Sauerland, 2009, Sauerland, 2010). In the case of ‘exactly one’-SAA, deriving an embedded UBC renders the sentence true; in the case of ‘every’-SAA, deriving an embedded UBC renders the sentence false. Hence, according to this constraint, and in line with our results, participants should derive embedded UBCs in the former case but not in the latter.

Again, however, this approach faces a number of problems. First, it fails to explain why participants readily judge sentences with unembedded ‘some’ false if their UBC is falsified. For example, Bott and Noveck (2004) found that participants judge sentences such as ‘Some dogs are mammals’ to be false about half the time (cf. Noveck, 2001). This is unexpected if hearers have a preference for truth. Similarly, the majority of participants still gave ‘false’ responses for ‘exactly one’-SAA and ‘no’-NAA, and, in the original study of Potts and colleagues, a substantial number of participants (27%) gave ‘false’ responses for ‘exactly one’-NNA. Both of these observations speak against the idea that hearers have a preference for truth, deriving UBCs only if doing so leads to a widening of the truth conditions.

The fate of the lexical uncertainty model proposed by Potts and colleagues seems tied to that of the other conventionalist accounts, as it predicts substantial rates of UBCs for ‘some’ in all embedding environments. However, the lexical uncertainty model might be sufficiently flexible to account for our data. Recall that, according to the lexical uncertainty model, the hearer determines, for each possible meaning of the words the speaker used, which sentences the speaker would have produced in which situations. Based on this information, the hearer constructs a conditional probability distribution from utterances to interpretations.

The predictions of their account depend on a number of factors. First, it is necessary to connect the conditional probabilities that the model predicts to actual rates of embedded UBCs. In the foregoing, we have assumed that the conditional probabilities should be normalised to the [0, 1] interval, but a different linking hypothesis would have entailed different predictions. Second, the predictions of the model depend on the set of alternative sentences the speaker could have uttered. Potts and colleagues assume a relatively idiosyncratic set of alternatives, consisting of sentences with ‘some’, ‘all’, and ‘none’ embedded under ‘every’, ‘exactly one’, and ‘no’. A different set of alternatives would lead to substantially different predictions. Third, the lexical uncertainty model includes a parameter that modulates how deeply the hearer reasons about the speaker’s beliefs and intentions. Different settings for this parameter lead to different predictions.

Hence, the lexical uncertainty model might be made compatible with the low rates of embedded UBCs in the squares-and-circles task. We must leave to future research the question whether the flexibility of the model can be exploited in a principled way to explain the fluctuating rates of embedded UBCs.

In summary, the challenge for conventionalist accounts is to formulate a conceptually grounded set of auxiliary assumptions that accounts for the distribution of embedded UBCs. Without such auxiliary assumptions, conventionalist accounts risk losing their “empirical bite”, to use a phrase from Geurts and Pouscoulous (2009, p. 24). That is, the claim that ‘some’ may or may not be read as ‘some but not all’, without carefully delineating under which circumstances such UBCs occur, simply cannot be falsified (Geurts & van Tiel, 2013, p. 10). Our results speak against several of the more prominent auxiliary assumptions that have been proposed in the recent literature.

Needless to say, conventionalist accounts can also invoke the explanations that we drew upon in defending a pragmatic perspective, e.g., the salience of the contrast between ‘some’ and ‘all’ and the inherently contrastive nature of non-monotonic quantifying expressions. However, if a substantial portion of the embedded UBCs that we observed can be attributed to such mechanisms, the remaining frequency of embedded UBCs might become so low as to call into question the conventionalist thesis that UBCs as a rule are the result of truth-conditional narrowing.

4.3. Experimental consequences

We began this paper by drawing an analogy between the conflicting findings in the Wason Selection Task and in sentence-picture verification tasks testing the prevalence of embedded UBCs. Our analogy was inspired by the idea that both debates include camps convinced that the sought inference—i.e., finding falsifying evidence of a conditional rule in the Wason Selection Task and the production of embedded UBCs in the sentence-picture verification tasks—is inevitable once the parameters of the task are set right.

One could argue that this analogy only has limited value because, whereas the literature on the Wason Selection Task developed tasks that rather quickly and reliably provided the sought normative responses, this does not seem to be the case with embedded UBCs. While we made some inroads here into uncovering features that allow participants to incrementally increase the drawing of UBCs, we remain unconvinced at this point that there are conditions under which they are routinely produced. This is a challenge for any conventional account that assumes that refined interpretations of 'some' are routine.

As one of our reviewers pointed out, there is another aspect in which the analogy between the debates about the Wason Selection Task and about embedded UBCs breaks down: in the latter debate, experimentalists very quickly started to introduce data from other, and in some cases novel, types of experimental tasks, whereas variations on the Wason Selection Task have tended to stay much closer to the original version (e.g., Oaksford, Chater, Grainger, & Larkin, 1997, Wason & Green, 1984). To illustrate, the following is a selection of the tasks other than the sentence-picture verification task that have been used to test how frequent embedded UBCs are:

i. *Inference task*. In their Exp. 1, Geurts and Pouscoulous (2009) simply asked participants whether they would infer from, e.g., 'All students heard some of the Verdi operas' that all students heard some but not all of the Verdi operas.

ii. *Graded truth judgement task*: Chemla and Spector (2011) presented participants with sentences and displays and asked them to indicate on a continuous scale how well the sentences described the corresponding displays.

iii. *Picture selection task*: Clifton and Dube (2010) presented participants with sentences and two displays and asked them to select which display was best described by the sentence. In addition, participants could indicate that the sentence described both displays equally well or neither.

iv. *Covered picture task*: Franke and colleagues (2017) presented participants with a sentence and a partially covered display and asked them to indicate whether they were able to determine the truth value of the sentence or whether they needed to uncover more of the display.

v. *Choosing actions task*: Benz and Gotzner (2017) presented participants with a sentence and asked them to indicate what actions would be licensed given the information expressed by the sentence.

vi. *Act-out task*: Tian, Breheny, and van Tiel (2012) presented participants with sentences such as 'Make sure that every set contains a ball in its boxes A and

B', which might lead to the embedded UBC that the hearer should make sure every set contains a ball *only* in its boxes A and B. They determined how participants complied with such instructions.

All of these experiments provide valuable data that need to be accounted for by any adequate theory of UBCs. At the same time, the experimental record is currently not amenable to a univocal interpretation, since it provides support for both pragmatic and conventionalist accounts. To provide an example, the 'exactly one'-NNSA condition in our squares-and-circles task was associated with an embedded UBC 12% of the time, whereas participants in Chemla and Spector's graded truth judgement task indicated that the corresponding sentence was 73% true in a similarly abstract NSA display.

What is needed, then, is a more thorough understanding of what drives participants in these tasks and, more generally, how these tasks relate to the various theories of UBCs. Such an understanding involves answering at least two questions. First, what are the response variables indicative of? For example, what does it mean if participants judge a sentence 73% true in a particular display (Chemla & Spector, 2011)? What does it mean if participants prefer a sentence to describe one display over another (Clifton & Dube, 2010). In our case, what does it mean if about 40% of the participants answer 'true' in the 'exactly one'-SAA condition? How do these response variables relate to mechanisms such as implicature, ambiguity, and typicality? Second, what methodological factors affect the behaviour of participants in these tasks? How is it possible, e.g., that, within the same experimental paradigm, one study finds evidence for a pragmatic account (Geurts & Pouscoulous, 2009) and another for a conventionalist account (Potts et al., 2016) of UBCs?

It lies outside of the scope of this paper to address these two questions for all of the experimental paradigms listed above. However, we have made a first step determining which methodological factors affect the prevalence of embedded UBCs within the sentence-picture verification paradigm, thereby explaining, at least in part, why sentence-picture verification tasks have been associated with markedly different results. We hope to have shown, first, that such a methodological approach may offer important insights into the mechanisms that underlie UBCs, and, second, that it is indispensable to use a sufficiently broad sample of materials when conducting research in pragmatics.

Returning to the Wason Selection Task, we should emphasise that we do not want to suggest that researchers on embedded UBCs should necessarily follow the path of researchers on the Wason Selection Task. It might well be that, since these two topics are of such a different nature, an entirely different experimental approach is warranted. However, we believe it is instructive, when considering the conflicting data on embedded UBCs, to learn from how researchers in adjacent fields have dealt with similarly conflicting results.

4.4. Outlook

Conflicting data are a staple of psychological research. Hence, it should come as no surprise to find conflicting data in experimental pragmatics, as well (e.g., the debate about eye movements and UBCs, cf., Huang & Snedeker 2009 vs. Grodner, Klein, Carbary, & Tanenhaus 2010, or the debate about the projection behaviour of presuppositions, cf. Chemla 2009 vs. Geurts & van Tiel 2016). It might be tempting to attribute such conflicting results to methodological flaws, such as an experimental task that is not sensitive enough to detect certain interpretations or materials that are not engaging enough for participants. Unless backed up with experimental data, however, such armchair criticisms should not be sufficient grounds for altogether discarding data from the debate. Indeed, these criticisms are more fruitful when viewed as sources of variability rather than reasons for discarding certain sets of results.

A much more interesting course of action—both from a theoretical and experimental perspective—is to determine which methodological features are responsible for conflicting findings, and what this tells us about the phenomenon in question (cf. Degen & Tanenhaus 2011, Gibson, Piantadosi, & Levy 2017, Phillips, Ong, Surtees, Xin, Williams, Saxe, & Frank 2015). This is the approach researchers on the Wason Selection Task have adopted, and we hope to have shown it is also the approach researchers in experimental pragmatics should take.

Acknowledgements

We would like to thank Jakub Dotlačil and all of the anonymous reviewers for their comments and questions. This research was supported by F.R.S.-FNRS Research Incentive Grant F.4502.15 (principal investigator M. Kissine) and by the German Research Council (grant DFG FR 3482/2-1, KR 951/14-1, SA 925/17-1) within SPP 1727 (Xprag.de), which are gratefully acknowledged.

BOB VAN TIEL
Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS)
bobvantiel@gmail.com

IRA NOVECK
Institut des Sciences Cognitives – Marc Jeannerod
ira.noveck@gmail.com

MIKHAIL KISSINE
Center of Research in Linguistics
Université Libre de Bruxelles
mkissine@ulb.ac.be

References

Abrams, K. H., Chiarello, C., Cress, K., Green, S., & Ellett, N. (1978). The relation between mother-to-child speech and word-order comprehension strategies in children. In R. N. Campbell, & P. T. Smith (Eds.) *Recent advances in the psychology of language. Vol. 4a: Language development and mother-child interaction*, (pp. 337–347). New York, NY: Plenum Press.

Ariel, M. (2015). Doubling up: two upper bounds for scalars. *Linguistics*, *53*(3), 561–610. http://dx.doi.org/10.1515/ling-2015-0013.

Banga, A., Heutinck, I., Berends, S. M., & Hendriks, P. (2009). Some implicatures reveal semantic differences. In B. Botma, & J. van Kampen (Eds.) *Linguistics in the Netherlands 2009*, (pp. 1–13). Amsterdam: John Benjamins.

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: the role of scalar alternatives in children's pragmatic inference. *Cognition*, *118*(1), 84–93. 10.1016/j.cognition.2010.10.010.

Benz, A., & Gotzner, N. (2014). Embedded implicatures revisited: issues with the truth value judgement paradigm. In J. Degen, M. Franke, & N. Goodman (Eds.) *Proceedings of the Formal & Experimental Pragmatics Workshop*, (pp. 1–6). Tübingen.

Benz, A., & Gotzner, N. (2017). Embedded disjunctions and the best response paradigm. To appear in: *Proceedings of Sinn und Bedeutung 21*.

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: the onset and time course of scalar inferences. *Journal of Memory and Language*, *51*(3), 437–457. http://dx.doi.org/10.1016/j.jml.2004.05.006.

Chemla, E. (2009). Presuppositions of quantified sentences: experimental data. *Natural Language Semantics*, *4*(17), 299–340. https://doi.org/10.1007/s11050-009-9043-9.

Chemla, E., & Spector, B. (2011). Experimental evidence for embedded implicatures. *Journal of Semantics*, *28*(3), 359–400. https://doi.org/10.1093/jos/ffq023.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, *17*, 391–416. https://doi.org/10.1016/0010-0285(85)90014-3.

Chierchia, G. (2004). Scalar implicatures, polarity phenomena and the syntax/pragmatics interface. In A. Belletti (Ed.) *Structures and beyond*, (pp. 39–103). Oxford: Oxford University Press.

Chierchia, G. (2006). Broaden your views: implicatures of domain widening and the 'logicality of language'. *Linguistic Inquiry*, *37*(4), 535–590. http://dx.doi.org/10.1162/ling.2006.37.4.535.

Chierchia, G., Fox, D., & Spector, B. (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In P. Portner, C. Maienborn, & K. von Heusinger (Eds.) *An international handbook of natural language meaning*, (pp. 2297–2332). Berlin: Mouton de Gruyter.

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*, 472–517. http://dx.doi.org/10.1016/0010-0285(72)90019-9.

Clifton, C., & Dube, C. (2010). Embedded implicatures observed: a comment on Geurts and Pouscoulous (2009). *Semantics and Pragmatics*, *3*(7), 1–13. https://doi.org/10.3765/sp.3.7.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.) *The adapted mind: evolutionary psychology and the generation of culture*, (pp. 163–228). New York: Oxford University Press.

Degen, J., & Tanenhaus, M. K. (2011). Making inferences: the case of scalar implicature processing. In L. Carlson, C. Hölscher, & T. Shipley (Eds.) *Proceedings of the 33rd annual conference of the Cognitive Science Society*, (pp. 3299–3304). Austin, TX: Cognitive Science Society.

Dehaene, S. (1997). *The number sense*. New York, NY: Oxford University Press.

Dieussaert, K., Verkerk, S., Gillard, E., & Schaeken, W. (2011). Some effort for some: further evidence that scalar implicatures are effortful. *The Quarterly Journal of Experimental Psychology*, *64*(12), 2352–2367. https://doi.org/10.1080/17470218.2011.588799.

Franke, M., Schlotterbeck, F., & Augurzky, P. (2017). Embedded scalars, preferred readings and prosody: an experimental revisit. *Journal of Semantics*, *34*(1), 153–199. http://dx.doi.org/10.1093/jos/ffw007.

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

Geurts, B. (2009). Scalar implicature and local pragmatics. *Mind & Language*, *24*(1), 51–79. https://doi.org/10.1111/j.1468-0017.2008.01353.x.

Geurts, B. (2010). *Quantity implicatures*. Cambridge: Cambridge University Press.

Geurts, B., & Pouscoulous, N. (2009). Embedded implicatures?!? *Semantics and Pragmatics*, *2*(4), 1–34. http://dx.doi.org/10.3765/sp.2.4.

Geurts, B., & van Tiel, B. (2013). Embedded scalars. *Semantics and Pragmatics*, *6*(9), 1–37. http://dx.doi.org/10.3765/sp.6.9.

Geurts, B., & van Tiel, B. (2016). When “all the five circles” are four: new exercises in domain restriction. *Topoi*, *35*(1), 109–122. http://dx.doi.org/10.1007/s11245-014-9293-0.

Gibson, E., Piantadosi, S. T., & Levy, R. (2017). Post hoc analysis decisions drive the reported reading time effects in Hackl, Koster-Hale & Varvoutis (2012). *Journal of Semantics*, *34*(3), 539–546. https://doi.org/10.1093/jos/ffx001.

Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. L. Morgan (Eds.) *Syntax and semantics, volume 3: Speech acts*, (pp. 41–58). New York: Academic Press.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology*, *73*(3), 407–420. https://doi.org/10.1111/j.2044-8295.1982.tb01823.x.

Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). "Some," and possibly all, scalar inferences are not delayed: evidence for immediate pragmatic enrichment. *Cognition*, *116*(1), 42–55. http://dx.doi.org/10.1016/j.cognition.2010.03.014.

Horn, L. R. (1972). *On the semantic properties of logical operators in English*. Ph.D. thesis, University of California, Los Angeles. Distributed by Indiana University Linguistics Club.

Horn, L. R. (2004). Implicature. In L. R. Horn, & G. Ward (Eds.) *The Handbook of Pragmatics*, (pp. 2–28). Malden, MA: Blackwell.

Horn, L. R. (2006). The border wars: a neo-Gricean perspective. In K. von Heusinger, & K. Turner (Eds.) *Where semantics meets pragmatics*, (pp. 21–48). Berlin: Mouton de Gruyter.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*(3), 346–363. http://dx.doi.org/10.1002/bimj.200810425.

Huang, Y. T., & Snedeker, J. (2009). On-line interpretation of scalar quantifiers: insight into the semantics-pragmatics interface. *Cognitive Psychology*, *58*(3), 376–415. http://dx.doi.org/10.1016/j.cogpsych.2008.09.001.

Inhelder, B., & Piaget, J. (1959). *La genèse des structures logiques élementaires: classifications et sériations*. Neuchâtel: Delachaux et Niestlé.

Johnson-Laird, P. N., Legrenzi, P., & Legrenzi, M. S. (1972). Reasoning and a sense of reality. *British Journal of Psychology*, *63*(3), 395–400. https://doi.org/10.1111/j.2044-8295.1972.tb01287.x.

Levinson, S. C. (2000). *Presumptive meanings: the theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Meyer, M.-C., & Sauerland, U. (2009). A pragmatic constraint on ambiguity detection: A rejoinder to Büring and Hartmann and to Reis. *Natural Language & Linguistic Theory*, *27*(1), 139–150. https://doi.org/10.1007/s11049-008-9060-2.

Noveck, I. A. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition*, *78*(2), 165–188. http://dx.doi.org/10.1016/s0010-0277(00)00114-1.

Noveck, I. A., & O'Brien, D. P. (1996). To what extent do pragmatic reasoning schemas affect performance on Wason's selection task? *Quarterly Journal of Experimental Psychology, Section A*, *49*(2), 463–489. https://doi.org/10.1080/027249896392739.

Noveck, I. A., & Sperber, D. (2007). The why and how of experimental pragmatics: the case of 'scalar inferences'. In N. Burton-Roberts (Ed.) *Advances in pragmatics*, (pp. 184–212). Basingstoke: Palgrave.

Nunberg, G. D. (1978). *The pragmatics of reference*. Ph.D. thesis, City University, New York. Distributed by Indiana University Linguistics Club.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: the probabilistic approach to human reasoning*. Oxford: Oxford University Press.

Oaksford, M., Chater, N., Grainger, B., & Larkin, J. (1997). Optimal data selection in the reduced array selection task (RAST). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(2), 441–458. http://dx.doi.org/10.1037/0278-7393.23.2.441.

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition*, *78*(3), 253–282. http://dx.doi.org/10.1016/s0010-0277(02)00179-8.

Phillips, J., Ong, D. C., Surtees, A. D. R., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A second look at automatic theory of mind. *Psychological Science*, *26*(9), 1353–1367. https://doi.org/10.1177/0956797614558717.

Potts, C., Lassiter, D., Levy, R., & Frank, M. C. (2016). Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics*, *33*(4), 755–802. https://doi.org/10.1093/jos/ffv012.

Pouscoulous, N., Noveck, I. A., Politzer, G., & Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Language Acquisition*, *14*(4), 347–375. http://dx.doi.org/10.1080/10489220701600457.

Sauerland, U. (2010). Embedded implicatures and experimental constraints: a reply to Geurts & Pouscoulous and Chemla. *Semantics and Pragmatics*, *3*(2), 1–13. https://doi.org/10.3765/sp.3.2.

Soames, S. (1982). How presuppositions are inherited: a solution to the projection problem. *Linguistic Inquiry*, *13*(3), 483–545.

Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition*, *57*(1), 31–95. https://doi.org/10.1016/0010-0277(95)00666-m.

Stenning, K., & van Lambalgen, M. (2004). A little logic goes a long way: basing experiment on semantic theory in the cognitive science of conditional reasoning. *Cognitive Science*, *28*(4), 481–530. https://doi.org/10.1016/j.cogsci.2004.02.002.

Storto, G., & Tanenhaus, M. K. (2005). Are scalar implicatures computed online? In E. Maier, C. Bary, & J. Huitink (Eds.) *Proceedings of Sinn und Bedeutung 9*, (pp. 431–445). Nijmegen: Nijmegen Centre for Semantics.

Tian, Y., Breheny, R., & van Tiel, B. (2012). Embedded implicatures: do they exist? Poster presented at *AMLaP* 2012, Riva del Garde, Italy.

van Tiel, B. (2014). Embedded scalars and typicality. *Journal of Semantics*, *31*(2), 147–177. http://dx.doi.org/10.1093/jos/fft002.

van Tiel, B., & Schaeken, W. (2017). Processing conversational implicatures: Alternatives and counterfactual reasoning. *Cognitive Science*, *41*(5), 1119–1154. http://dx.doi.org/10.1111/cogs.12362.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140. https://doi.org/10.1080/17470216008416717.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. https://doi.org/10.1080/14640746808400161.

Wason, P. C., & Green, D. W. (1984). Reasoning and mental representation. *The Quarterly Journal of Experimental Psychology Section A*, *36*(4), 597–610. https://doi.org/10.1080/14640748408402181.

Wason, P. C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology*, *23*(1), 63–71. https://doi.org/10.1080/00335557143000068.

Appendix: Filler items

The following 13 filler items were used in Exp. 1: Squares-and-circles and Exp. 3.

(12) a. The red circle is connected.
b. The red and green squares are connected to the same number of circles.
c. The blue square is not connected.
d. The red circle is connected to the most squares.
e. The green circle is connected.
f. The green and blue circles are connected to the same number of squares.
g. The red circle is the second most connected shape.
h. The blue circle is not connected.
i. The green square is connected to the fewest circles.
j. There is a shape that has more connections than the green square.
k. All the circles are connected to the same number of squares.
l. Most of the squares are connected to circles.
m. One of the circles has the fewest connections.

The following 13 filler items were used in Exp. 1: Basketball, Exp. 2, Exp. 4, Exp. 5, and Exp. 6.

(13) a. Player A shot perfectly.
b. Player A tied with Player B.
c. Player A made no shots.
d. Player B was the star player of the round.
e. Player B shot perfectly.
f. Player B tied with Player C.
g. Player C placed second.
h. Player C messed up.
i. Player C placed last.
j. Someone did better than Player C.
k. All the players are tied.
l. Most of the players missed shots.
m. We have a clear loser.

The L+ tasks in Exp. 2, in addition, included the following 10 (C+) or 16 (C–) filler items.

(14) a. Player A was the star player of the round.

b. Player A placed second.
c. Player A tied with Player C.
d. Player A placed last.
e. Player A messed up.
f. Player A made none of his baskets.
g. Someone did better than Player A.
h. Player B placed second.
i. Player B made no shots.
j. Player B placed last.
k. Player B messed up.
l. Someone did better than Player B.
m. Player C shot perfectly.
n. Player C made no shots.
o. Most of the players made shots.
p. We have a clear winner.