Where truth and optimality part. Experiments on implicatures with epistemic adverbs

Adina Camelia Bleotu, Anton Benz & Nicole Gotzner*

Abstract. In the current paper, we employ a novel Shadow Play Paradigm in order to test Romanian monolingual adults' sensitivity to truth and informativeness and investigate their ability to derive implicatures with epistemic adverbs. We show that implicature rates with epistemic adverbs are higher when participants are asked to reward characters depending on the truth of their statements rather than on whether what they say is the best description of the situation. Given participants' task-sensitivity, we recommend that instructions use optimality criteria, as they are a more sensitive method of probing into implicature generation.

Keywords. scalar implicatures; modality; epistemic adverbs; Truth Value Judgment Task; Optimality Judgment Task; Romanian; methodology

1. Introduction. Experimental paradigms that study pragmatic reasoning by either testing the fit of sentences to situations or the fit of situations to sentences can be divided into two broad groups: They either ask subjects to make judgements about truth and falsity, or they ask them to make judgements about some measure of appropriateness. We argue that paradigms employing judgements about truth and falsity activate reasoning about semantic meaning, while judgments about appropriateness activate reasoning about pragmatic meaning. We consider this hypothesis in the context of a case study that investigates the derivation of scalar implicatures with the epistemic adverb *poate* 'maybe' in Romanian in the case of Romanian monolingual adults by means of a novel Shadow Play Paradigm. We implement two reward versions of this paradigm, one that asks subjects to reward characters based on judgements about truth and falsity (Right-Wrong task) and one that asks them to reward characters based on judgements about the appropriateness of sentences (Best Description Task). Our experiments show that implicature rates with the epistemic adverb *poate* 'maybe' are significantly higher in the case of the optimality judgment task (Best Description Task) than in the Truth Value Judgment Task (Right-Wrong Task), thus emphasizing an important methodological point: that results and, consequently, the theory accounting for them are largely dependent upon the methods used.

The paper is organized as follows: After a brief introduction, in Section 2, we present previous research on scalar implicatures from a methodological perspective. Section 3 deals with previous research on epistemic modality in language acquisition. Section 4 describes the experiments we conducted (Goals, Participants, Methodology, Results). Section 5 discusses the results and their consequences for future research on scales. Section 6 draws a conclusion on this basis.

* This research was supported by an XPrag.de internship offered to Adina Camelia Bleotu at ZAS Berlin within the Deutsche Forschungsgemeinschaft (DFG) project *SI Games I: Experimental Game Theory and Scalar Implicatures* led by Dr. Anton Benz (Grant Nr.: BE 4348/4-1). Anton Benz was supported by the Bundesministerium für Bildung und Forschung (BMBF), Grant Nr. 01UG1411. Nicole Gotzner was supported by the DFG, Grant Nr. BE 4348/4-2, through the priority program *New Pragmatic Theories based on Experimental Evidence* (SPP 1727), and she is further supported by the DFG through the Emmy Noether Programme (Grant Nr. GO 3378/1-1). We are grateful to the students from the Faculty of Foreign Languages and Literatures, University of Bucharest, who took part in the experiments. Authors: Adina Camelia Bleotu, ICUB, University of Bucharest (cameliableotu@gmail.com), Anton Benz, ZAS Berlin (benz@leibniz-zas.de), & Nicole Gotzner, ZAS Berlin (gotzner@leibniz-zas.de) and University of Potsdam.

©2021 Adina Camelia Bleotu, Anton Benz and Nicole Gotzner
Published by the LSA with permission of the author(s) under a CC BY license.

2. Previous research on scalar implicatures. A methodological perspective. An important methodological decision is how to find out whether participants derive implicatures or not (Noveck 2001, Papafragou & Musolino 2003, Geurts & Pouscolous 2009, Clifton & Dube 2010, Katsos & Bishop 2011, van Tiel 2013, Benz & Gotzner 2014, a.o.). Experiments are essentially classified into whether they test the fit of sentences to situations (the most frequent method), the fit of situations to sentences, or whether they resort to some indirect measure of the responses given by participants (such as rewards in the Best Response Paradigm in Gotzner & Benz 2018). Given that the literature on implicatures is extremely vast, and covering it is beyond the scope of this paper, our presentation will mainly focus on some relevant examples of the most commonly used paradigm, the paradigm which tests the fit of sentences to situations.

In order to find out whether subjects derive implicatures when confronted with a statement like *Some dogs are black/The dogs may be behind the curtain* in a context where all dogs are black/behind the curtain, experimental paradigms initially employed questions used in the original Truth Value Judgment Task (Crain & McKee 1985, Crain & Thornton 1998) such as *Is the puppet right?* or *Do you agree with the statement?* (Noveck 2001). However, Papafragou & Musolino (2003:264) challenged this methodology, arguing that such questions actually tap into speakers' sensitivity to truth:

> "In our version, instead of asking subjects if the puppet is 'right' or 'wrong' (as in the original TVJT), we asked whether the puppet 'answered well' (i.e., *Apantise kala*, 'Did-(she)-answer well?'). This modification was made since we were interested in felicity, not truth."

While running a Truth Value Judgment Task calls for questions about right/wrong or true/false, agree/disagree, running a Felicity Judgment Task calls for questions about adequacy/appropriateness. This idea has been explored in various ways in the literature dealing with scalar implicatures, which consequently experienced a methodological shift from truth-oriented to felicity-oriented tasks (Katsos & Bishop 2011). In a sense, even asking if a puppet answered well may be too weak, given that subjects may assess certain underinformative statements as good enough, and, thus, still evaluate matters in terms of truth value. For this reason, experimental pragmatics has been trying to employ novel methods which make subjects sensitive to the difference between optimal statements (the best, most felicitous statements in a certain pragmatic context) and statements that are true but less optimal.

Among the paradigms testing the fit of sentences to situations, one way of testing optimality rather than truth is by asking subjects to provide graded judgments. Katsos & Bishop (2011) tested 6- to 7-year-old English-speaking children for underinformativeness with existential quantifiers by means of a ternary Reward Task where children were asked to offer a 'small', 'big', or 'huge' strawberry as a reward to Mr. Caveman depending on how good the speaker's responses were. The paradigm revealed sensitivity to underinformativeness on the part of children, who rewarded such statements with big strawberries instead of huge or small ones. Given children's general acceptance of underinformative statements in a standard Truth Value Judgment Task (Katsos & Bishop 2011), the results from the ternary task were quite surprising, revealing that children had more pragmatic sensitivity than previously thought. This led Katsos & Bishop (2011) to argue that the yes-no binary task was not fine-grained enough to capture the difference between truth and informativeness, and, thus, it gave the illusion that children were insensitive to violations of informativeness, when, in fact, they were merely tolerant. A different manner of implementing graded judgments is by asking participants to rate certain sentences as descriptions of certain

images by placing the cursor between yes and no. Using the cursor placement method, Chemla & Spector (2011) conducted several experiments testing whether adults derive local implicatures in sentences like *Every letter is connected to some of its circles* ('Every letter is connected to some, not all of its circles.') and concluded that local implicatures are attested, against Geurts & Pouscoulous (2009).

Another way of testing pragmatic adequacy/optimality is the Felicity Judgment Task, which also tests the fit of sentences to situations, but through a Forced Choice Task where participants choose the best sentence out of two sentences. Foppolo, Guasti & Chierchia (2012) conducted a Felicity Judgment Task (Experiment 5 in the paper) on 5-year-old Italian children in order to investigate their ability to generate implicatures with existential quantifiers. The set-up involved two puppets who were always quarrelling about which of them described various pictures best, and children had to be the judge. For instance, Puppet 1 would utter a weak, underinformative statement with *qualche* 'some', while Puppet 2 would utter a stronger, informative statement with *tutti* 'all'. Then, children were asked: "Which puppet said it better?". A similar methodology was used by Ozturk & Papafragou (2015) in order to test children's ability to derive implicatures with epistemic *may*: children had to choose between the statements produced by Minnie and Donald about the location of an animal in one of two boxes (an underinformative statement with *may* and a fully informative statement with *have to*). Importantly, the tasks reveal adult-like answers on the part of children, showing that access to stronger alternatives eases pragmatic understanding.

There are also paradigms that do not ask for judgments about fit of sentences to situations, or situations to sentences, such as the Best Response Paradigm proposed by Gotzner and Benz (2018). An important point about the paradigm is that it avoids meta-linguistic judgments about either truth or appropriateness. Instead, judgments are read off from the rewards given by participants in an interactive game–theoretic reward task set-up which satisfies Grice's conversational requirements for implicature generation (a recognizable purpose for the talk exchange). In a task where four girls have lost all/some/none of their marbles and they have to find them again, adults have to handle an explicit decision problem, rewarding characters based on statements about how many marbles they find: (i) chocolate for finding all of the marbles, (ii) candy for fewer than all and (iii) a gummy bear for none of the marbles (as a consolation prize). Such a task led to many local implicatures in utterances like *All of the girls found some of their marbles* ('All of the girls found some, not all of their marbles'), revealing subjects' sensitivity to subtle considerations of informativeness. This paradigm revealed quite different results from binary Truth Value Judgment tasks (see also Benz & Gotzner, in press, for an interactive version of the Best Response Paradigm).

Apart from the notion of felicity, another notion that has been brought under discussion in relation to the adequacy of an underinformative utterance to a pragmatic context is typicality, that is, the degree to which the situation described by the utterance is a typical one (van Tiel 2013). Typicality has been studied experimentally starting with Rosch (1975), who asked participants to produce typicality orderings by evaluating hyponyms of BIRD (i.e., *robin*) along a 7-point Likert-scale. Typical members are learnt earlier, recognized faster and more accurately, and produced earlier (van Tiel 2013). Interestingly, many experiments on implicatures (Geurts & Pouscoulous 2009, Clifton & Dube 2010, Chemla & Spector 2011) are quite similar to Rosch's rating task: participants see a category with several instances/a sentence with several situations and have to decide how well the category/the sentence describes them. For this reason, van Tiel (2013) argues that typicality differences may influence the interpretation not only of predicates like *bird* but also of a quantifier like *some*. For instance, Begg (1987) found that the typical meaning of *some* is less than half, and Degen & Tanenhaus (2011) obtained similar results. In the context of associating

pictures representing various situations with statements, typicality may overlap with felicity or optimality, given that the best statement describing a situation is probably the most typical one as well- though various aspects seem to impact typicality (e.g. plausibility, world knowledge, a.o.).We note that for our purposes, in the case of epistemic items, both possibility and certainty items are licensed in contexts where one does not have direct access to the objects/animals under discussion. However, using *possible* instead of *certain* in a context of certainty would count as not typical, as well as infelicitous and not optimal, given that *possible* is expected only in uncertainty contexts. The previously noted advantages of methods relying on optimality/felicity lie at the basis of the current experiments on implicatures with epistemic adverbs, contrasting truth sensitivity to pragmatic sensitivity.

3. Previous research on epistemic modality. Methodological insights from language acquisition. Many experimental studies on epistemic modality (Hirst & Weil 1982, Noveck, Ho & Sera 1996, Noveck 2001, Ozturk & Papafragou 2015 a.o.) have focused on the acquisition of epistemic modal verbs. Such studies have shown that children are sensitive to the relative strength of modal verbs from very early on. Although aware of the existence of a modal scale, children still have difficulties with modals at age 5, achieving epistemic maturity only later on, around age 7.

The paradigm used in the studies on epistemic modality is some version of the Hidden Object Task, where, based on evidence, subjects have to infer the location of a certain hidden object. A first version of the task is represented by the *Look for the Peanut* Task (Hirst & Weil 1982), where children were asked to look for a peanut on the basis of certain statements with modals they heard. Noveck, Ho & Sera (1996) and Noveck (2001) then tested epistemic items by employing the Box Paradigm, where objects are hidden in boxes, a paradigm which was later on simplified by Ozturk & Papafragou (2015). In the initial box paradigm, there were three boxes (two uncovered boxes which contained an animal or two and a covered one), and subjects were interrogated about a third covered box based on disjunctive statements of the type "This box has the same content as either Box A or as Box B". Ozturk & Papafragou (2015) reduced the complexity of the paradigm by resorting to only two boxes and one single animal and by resorting to non-disjunctive input. The semantic Forced Choice Task conducted by Noveck, Ho & Sera (1996) and the semantic Truth Value Judgment Task conducted by Ozturk & Papafragou (2015) both showed that young children have a tendency to reduce uncertainty and accept situations where a stronger statement (with *has to*) is made instead of a weaker one (with *may*). In terms of pragmatics, Noveck (2001) conducted a Truth Value Judgment Task, where subjects had to say whether they agreed with a certain statement, while Ozturk & Papafragou (2015) ran a Felicity Judgment Task, where subjects had to choose between two statements (an underinformative statement versus a fully informative statement). 5-year-olds performed more adult-like in the Felicity Judgment Task than in the Truth Value Judgment Task, which further reinforces the idea that implicatures are more easily accessed by tasks focusing on pragmatic adequacy. Interestingly, regardless of the task type, the implicature rates with modals were quite high for adults (close to 90%). Nevertheless, in an adaptation of Noveck (2001) on epistemic adverbs in Romanian (Bleotu 2019), many adults were too cautious, rejecting statements about the certainty of something they could not see.

4. Current experiments: Truth and optimality in the Shadow Play Paradigm.

4.1. RATIONALE AND GOALS. Given subjects' caution in the Hidden Object Paradigm, i.e., in situations of no direct access to the object, we developed a novel Shadow Play Paradigm, where subjects have to reward a dragon for the statements he makes about the identity of a shadow/silhouette, on the basis of certain evidence. Importantly, unlike in the Hidden Object

Paradigm, in the Shadow Play Paradigm, subjects can infer that the shadow must belong to an animal by looking at its silhouette and also because of sounds that accompany it (e.g., *woof-woof*, representing a dog). By making indirect evidence more direct and, thus, making the task more evidential, we aimed to prevent subjects from being overly cautious and giving unreliable answers.

The paradigm itself draws inspiration from shadow play theater, an ancient art form with silhouettes. A similar but somewhat simpler paradigm was also employed by Heizmann (2006) in order to test whether children make indirect inferences with necessity modals in English and German. Heizmann (2006) used real characters and their silhouettes in order to see how children interpret questions such as *Who must be eating a banana?* versus *Who must eat a banana?*. The Reward Task was inspired from Katsos & Bishop (2011) but, instead of using a ternary Reward system, we used a binary reward system associated with different linguistic input.

In order to test whether subjects are more sensitive to underinformativeness than to truth value in deriving scalar implicatures, we decided to run the same shadow play test in two different versions: (1) a Right-Wrong Task, where subjects were asked to reward a baby dragon with a big/small apple depending on the truth value of his statement, and (2) a Best Description Task, where subjects had to reward a baby dragon with a big/small apple depending on whether what he said was the best description of the situation or not. The expectation was that participants would derive more implicatures in the Best Description Task, which encourages pragmatic sensitivity.

4.2. METHODOLOGY.

4.2.1. PARTICIPANTS. The Right-Wrong Task was conducted on 64 native Romanian speakers, and the optimality test was conducted on 63 Romanian native speakers, recruited from 1st and 2nd year students at the Faculty of Foreign Languages, University of Bucharest.

4.2.2. MATERIALS. We implemented two versions of our experiment in PennController (Zehr & Schwarz 2018). While the experiments employ the same type of task (a Reward Task), the criteria for rewarding were different: truth value ("right-wrong") (in the Right-Wrong Task) and optimality ("best description") (in the Best Description Task). The set-up was exactly the same for both experiment versions. The scenario was that of a *Shadow Play Paradigm*, telling participants that there is a wizard who likes to play the shadow game with a baby dragon. In this game, various animals go and hide behind the curtain—but some of them may come in front of the curtain later on. The baby dragon has to say who he thinks the shadow belongs to. Participants are told that they are supposed to reward the baby dragon with a big apple if what he says is right (Right-Wrong Task) / the best description (Best Description Task) and with a small apple if what he says is wrong (Right-Wrong Task) / not the best description (Best Description Task). Importantly, such a contrastive experimental set-up assumes that, in the Right-Wrong Task, subjects will reward both fully informative and underinformative true statements with a big apple, whereas, in the Best Description Task, they will only reward fully informative statements with a big apple, and underinformative true statements will receive a small apple, just like false ones (see Table 1):

RIGHT		WRONG
FULLY INFORMATIVE	UNDERINFORMATIVE	FALSE
THE BEST DESCRIPTION	NOT THE BEST DESCRIPTION	

Table 1: Truth, informativity and optimality

The experimental materials involve several associated pictures and sentences. Each picture has a main silhouette, a small image with the animals in front of the curtain, and a small image with all

the animals in the game (see Figures 1, 2). The small image on the left (ALL ANIMALS) is always present for subjects to easily access the initial situation, without processing difficulties because of memory load (Crain & Thornton 1998). There are various groups of animals of various colors: a training group of two bunnies and 4 testing groups of three animals each: dog, frogs, cats, cows.

We presented participants with 59 sentences in total (3 training sentences, 2x4=8 test sentences, 2x24=48 control sentences containing *poate* 'maybe' or *sigur* 'certainly' presented in a randomized manner, see Table 2). The randomization was applied both within the same group of animals and across groups. The test contains a number of sentences balanced between the two epistemic adverbs so as to activate the modal scale <*possible, certain*> and trigger pragmatic readings. The key sentences for implicature detection are highlighted in yellow. All the sentences (except for the practice ones) have the same structure: the adverb *poate* 'maybe' / *sigur* 'certainly' followed by the complementizer *că* 'that' and an embedded sentence.

NONE IN FRONT SCENARIO				ONE IN FRONT SCENARIO					
SPOSSIBLE1 UNDERINFO	SCERTAIN1 OPTIMAL	SPOSSIBLE2 FALSE	SCERTAIN2 FALSE	SPOSSIBLE3 OPTIMAL	SPOSSIBLE4 OPTIMAL	SCERTAIN3 OVERLY STRONG	SCERTAIN4 OVERLY STRONG	SPOSSIBLE5 FALSE	SCERTAIN5 FALSE
TWO IN FRONT SCENARIO									
SPOSSIBLE6 UNDERINFO	SCERTAIN6 OPTIMAL	SPOSSIBLE7 FALSE	SCERTAIN7 FALSE						

Table 2: Types of utterances tested per Scenario

4.2.3. PROCEDURE. The experiment started with a training session, followed by the main experiment. In the training session, participants get acquainted with the picture design and practice rewarding on the basis of a bunny shadow picture (see Figure 1), where they have to reward the baby dragon with big or small apples. Subjects were presented with sentences such as the ones in (1). In the first sentence in (1a), subjects were told which reward to choose (the small apple), while, in the other sentences, they had to choose the reward themselves. The training items were the same in both the Right-Wrong Task and in the Best Description Task.

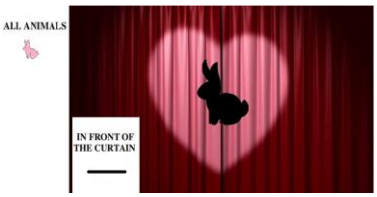

Figure 1: Item from training session

(1) a. Este un şoarece /o vacă. (FALSE)
'It is a mouse/a cow.'
b. Este un iepuraş. (TRUE/OPTIMAL)
'It is a bunny.'

We will now exemplify the testing session by reference to the group of dogs.

Scenario 1*, *the None in front Scenario (where all dogs go behind the curtain, see Figure 2) ensures that subjects have in mind the set of animals (*the referential domain*) that is *at issue*, rather than all the animals in the world, or the animals in the game. Sentences (2a) and (2b) are the critical

conditions. Sentence (2c) was a control sentence. Subjects are supposed to reason that the animal whose silhouette they see must be a dog, not a cat, not a cow, not a frog, and, hence, choose a small apple as reward. Subjects who only consider semantic meaning are expected to reward the dragon with a big apple in conditions (2a) and (2b), and subjects that strengthen the weak epistemic to 'not certain' are expected to choose a small apple for (2a) but not for (2b).

Figure 2: Example picture for the None/One/Two in front Scenarios

(2) a. Poate că este un câine. (UNDERINFO)
'It is possible that it is a dog.'
b. Sigur că este un câine (OPTIMAL)
'It is certain that it is a dog.'
c. Poate/sigur că este o pisică. (FALSE)
'It is possible/certain that it is a cat.'

Scenario 2, the ***One in front Scenario*** (where one animal comes back in front of the curtain, in this case, the yellow dog, see Figure 2) tests the subjects' understanding of alternatives, their ability to reason that the situation has two possible outcomes: either the silhouette belongs to the red dog, or it belongs to the blue dog. Subjects were expected to choose a big apple for the optimal control statements in (3a) and a small apple for the wrong statements in (3b, c).

(3) a. Poate că este câinele roşu/albastru. (OPTIMAL)
'It is possible that it is the red/blue dog.'
b. Sigur că este câinele roşu/albastru. (OVERLY STRONG)
'It is certain that it is the red/blue dog.'
c. Poate/Sigur că este câinele galben. (FALSE)
'It is possible/certain that it is the yellow dog.'

Scenario 3, the ***Two in front Scenario*** (where two animals are in front of the curtain, see Figure 2) tests whether subjects are able to reason that the silhouette can only belong to the blue dog, given that there are two animals in front of the curtain now. Sentences (4c, d) were control sentences. Subjects who only consider semantic meaning are expected to reward the dragon with a big apple in both (4a) and (4b), and subjects that strengthen the weak epistemic to 'not certain' are expected to choose a big apple for (4b) but not for (4a).

(4) a. Poate că este câinele albastru. (UNDERINFO)
'It is possible that it is the blue dog.'
b. Sigur că este câinele albastru. (OPTIMAL)
'It is certain that it is the blue dog.'
c. Poate că este câinele roşu. (FALSE)
'It is possible that it is the red dog.'
d. Sigur că este câinele roşu. (FALSE)
'It is certain that it is the red dog.'

4.3. RESULTS. All subjects were included in the analysis of the data given that there were no subjects who made more than 1 error out of 3 in the bunny control items. In order to see the difference in choices between the Right-Wrong Task and the Best Description Task, we performed analyses: (a) on the whole set of data, (b) on subsets of the data: (i) underinformative statements, (ii) (true optimal and false) control statements, and (iii) overly strong statements.

4.3.1. WHOLE DATA ANALYSIS. Using R (2018), We computed a logit mixed-effects model with Task type and Statement type (control, underinformative, overly strong) as fixed effects with treatment coding and Item and Participant as random effects. The model took the answers to the control condition as the baseline. The results show that the type of task, the overly strong statement type and the underinformative statement type are statistically significant (see Table 3). In addition, the interaction between the type of task and the underinformative condition is also significant, but not the interaction between the type of task and the overly strong statement type.

Parameter	Estimate	Std. error	z	p
Intercept	2.794	0.227	12.293	< 2e-16 ***
Task type right wrong	-0.893	0.193	-4.634	3.58e-06 ***
Condition Overly Strong	-1.033	0.149	-6.931	4.17e-12 ***
Condition Underinformative	-1.423	0.14	-10.096	< 2e-16 ***
Task type right wrong: Condition Overly Strong	0.097	0.182	0.536	0.592
Task type right wrong: Condition Underinformative	-0.913	0.178	-5.121	3.03e-07

Table 2: Results of a glmer performed on the whole data

4.3.2. SUBSETS OF THE DATA. We divide the subset analysis into three parts: scalar implicatures, control statements, and overly strong statements.

For scalar implicatures, to determine rates of implicature with more precision, we looked at the corresponding stronger alternative statements with *sigur* 'certainly'. Subjects were assumed to derive scalar implicatures when they gave a small apple reward to the underinformative statement with *poate* 'maybe' (2a, 4a) and a big apple reward to the stronger alternative statement with *sigur* 'certainly'. Interestingly, whereas in the Right-Wrong Task, only 29.24% speakers rejected underinformative sentences, with only 14 consistent speakers (i.e., giving more than 5 expected answers out of 8), in the Best Description Task, there were 66.67% scalar answers, with 41 consistent subjects (see Figure 3). We ran a logistic regression using a logit mixed-effects model with Scalar implicatures as variable, Task type and Scenario as fixed effects and Item and Participant as random effects. The results reveal a significant effect for Task (β = −3.503, SE = 1.664, Z = −5.274, p < 0.001) and the interaction between Task and Scenario (β = 0.393, SE = 0.196, Z = 2.011, p = 0.044), but no significant effect per Scenario (β = −0.203, SE = 0.1435, Z = −1.417, p = 0.156).

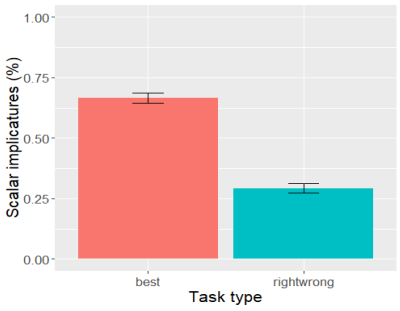

Figure 3: Scalar implicatures per task

The results for control sentences (see Table 3) reveal more accuracy in the Best Description Task than in the Right-Wrong Task, indicating that the Best description encourages participants' attention. We conducted a logistic regression using a mixed-effects model with Task and Truth as fixed effects and random by-item and by-participant slopes. The results reveal a significant effect per Task (β = −1.164, *SE* = 0.199, *Z* = −5.841, $p < 0.001$), Truth (β = −0.904, *SE* = 0.105, *Z* = −8.56, $p < 0.001$), as well as the interaction between Task and truth (β = 0.265, *SE* = 0.129, *Z* = 2.052, $p = 0.004$).

Accuracy	Best Task	Right-Wrong Task
Optimal control sentences	84.02%	78.83%
False control sentences	92.27%	80.95%

Table 3: Accuracy in control sentences per task

Interestingly, in the case of overly strong statements, there was also more accuracy with the Best Description Task (35.67%) than with the Right-Wrong Task (22.17%), see Figure 6. A logistic regression using a mixed-effects model with Task as a fixed effect and Item and Participant as random effects reveals a significant task effect (β = −2.877, *SE* = 0.642, *Z* = −4.482, $p < 0.001$), while including random by-item and by-participant slopes leads to near significance per Task (β = 2.607, *SE*= 2.607, *Z* = 1.914, $p = 0.055$).

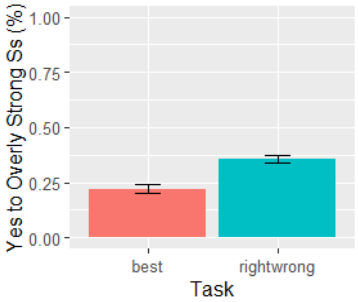

Figure 6: Yes to overly strong sentences per task

5. Discussion. The whole set analysis indicates that participants behave more accurately with underinformative statements in the Best Description Task. However, we decided to also do subset analyses of the data, given that awarding the dragon with a small apple for underinformative statements is not necessarily an indication of scalar implicatures. To see this, note that there are at least two possible reasons for rejecting the underinformative sentence (4a), either (a) the participant thinks it is impossible for the silhouette to be the blue dog, in which case he/she would also reject the stronger alternative, or (b) the subject thinks it is actually certain, not just possible that it is the blue dog, in which case he/she would accept 'It is certain that it is the blue dog'. Since we did not ask subjects why they gave certain answers-in order to keep the test relatively short, both options are possible in the current experimental set-up. Thus, we decided to analyze whether participants derive implicatures by looking at the underinformative sentences (in both scenarios) and at the corresponding stronger alternative statements with *certain(ly)*. We noticed four different patterns of responses in our participants (see Table 4): a logical pattern, for subjects who accepted both optimal and underinformative statements, a pragmatic pattern, for subjects who rejected the underinformative statement, but accepted the optimal one, a cautious pattern, for subjects who accepted the underinformative statements, but not the optimal ones with *sigur* 'certainly', and an erroneous pattern, for subjects who rejected both the underinformative and optimal statements.

It is possible that it is X	It is certain that it is X	Pattern of response	Best Task	Right-Wrong Task
big apple	big apple	Logical	24.6%	46%
small apple	big apple	Pragmatic	66.67%	29.24%
big apple	small apple	Cautious	1.38%	11.7%
small apple	small apple	Erroneous	7.34%	13.06%

Table 4: Subjects' patterns of response in underinformative and optimal true sentences

Interestingly, there is no difference between the proportions of scalar implicatures derived with underinformative sentences belonging to the None in front Scenario or the Two in front Scenario. Evaluating whether the silhouette belongs to a certain animal class or is an animal of a certain color leads to similar results (see Table 5), suggesting pragmatic consistency across scenarios.

Patterns of responses	None in front Scenario		Two in front Scenario	
	Best Task	Right-Wrong Task	Best Task	Right-Wrong Task
Pragmatic	68.65%	27.34%	64.68%	31.25%
Logical	27.38%	53.9%	21.28%	38.28%
Cautious	1.19%	5.07%	1.58%	18.36%
Erroneous	2.77%	14%	11.9%	12.1%

Table 5: Subjects' patterns of responses per task and scenario

The results from our Right-Wrong Task with epistemic adverbs reveal quite low rates of implicatures (29.24%). In contrast, the Best Description Task leads to higher implicature rates (66.67%). On the one hand, it is unclear whether the low rates in the Right-Wrong Task could be explained through the more complex nature of the scalar items tested (epistemic adverbs), as well as language-specific facts related to Romanian (the fact that epistemic adverbs select full CPs, for instance). We would expect these factors to affect implicature-derivation in the Best Description Task as well, not just in the Right-Wrong Task. On the other hand, both the Best Description task and the Right-Wrong Task were implemented as a binary task, so the results cannot be understood in terms of an opposition between binary and ternary (as in Katsos & Bishop 2011). Rather, the essential aspect seems to be related to how task instructions model participants' attention: the Right-Wrong Task encourages adults to pay attention to the truth value of the statements they hear, whereas the Best Description Task encourages adults to pay attention to informativity.

As far as the control statements are concerned, the results again suggest better accuracy with the Best Description Task. Importantly, the fact that participants had lower accuracy on the true statements suggests there was no yes bias, but rather a tendency to place 'bets' on certain animals and reject statements about the possible presence of other animals behind the curtain.

In the case of overly strong statements, both tasks had participants who rewarded dragons with big apples (using *sigur* 'certainly' where the weaker *poate* 'maybe' was optimal). The quite high number of big apple rewards for overly strong statements is unexpected given that such statements are false, not optimal. While this could be due to inattention, the higher accuracy in the control statements sheds doubt upon such an explanation. We believe that such answers actually reflect a tendency to place a 'bet' on one of the animals when it is yet unknown what animal lies behind the curtain, a tendency which is encouraged by the present task. The Best Description Task taps into subjects' awareness that they should not place a bet on a certain outcome, hoping that it is true, but rather evaluate whether the statement they hear corresponds to the situation in the best

way possible or not. Thinking whether the baby dragon described the situation best makes it clearer that they are not dealing with a guessing game, but rather a reward game based on evidence.

6. Conclusion. The results show a significant task effect in the derivation of scalar implicatures with *poate* 'maybe' when comparing the Truth Value Judgment (the Right-Wrong Task) to the Optimality Judgment (the Best Description Task). Interestingly, accuracy seems to be better in the Best Description Task even with control statements and overly strong statements, which suggests that asking adults to reward characters depending on whether their statement is the best description or not makes adults more attentive. In contrast, there may be more tolerance with respect to what they consider right and wrong. This is very much in line with the Katsos & Bishop (2011) language acquisition findings. Katsos & Bishop (2011) argue that the Right-Wrong Binary Task masks children's sensitivity to underinformativeness due to their pragmatic tolerance. In other words, children tend to consider underinformative statements true, but they realize underinformative statements are not optimal. This becomes obvious in their ternary task, where children reward optimal, underinformative, and false statements differently.

In the Right-Wrong Task, there were low rates of implicatures, whereas, in the Best Description Task, there were high implicature rates. This indicates that adults are sensitive to task instructions. Importantly, testing adults' interpretation of the stronger alternatives to the underinformative statements allows us to make an informed decision about whether speakers derived scalar implicatures or not. In this way, we can evaluate not only adults' sensitivity to underinformativeness in the Best Description Task, but their actual ability to derive implicatures.

The current experiments thus show the importance of methodology in research: (even linguistically naïve) adults generate implicatures only when asked the adequate question. For these reasons, we recommend using test questions about optimality, especially considering previous research where inferences were not robust. Nevertheless, an important point is in order: deciding whether participants derive implicatures implies establishing whether they consider a certain statement true yet underinformative. While the Right-Wrong Task masks sensitivity to underinformativeness (since the statements rewarded with big apples are either optimal or underinformative), the Best Description Task, which was implemented as a binary task as well, masks adults' truth evaluations (since the statements rewarded with small apples are either underinformative or false). Hence, it is extremely important to either ask questions about participants' reasons for giving a certain answer or include control sentences that evaluate whether participants accepted the stronger alternatives (of underinformative statements)-the latter represents the strategy we adopted in our experiments. Another option is to resort to ternary tasks, which test participants' understanding of optimal, underinformative, and false statements at once, through a three-valued reward system.

References

Begg, Ian. 1987. 'Some'. *Canadian Journal of Psychology* 41:62–73. https://doi.org/10.1037/h0084147.

Benz, Anton & Nicole Gotzner. 2014. Embedded implicatures revisited: Issues with the truth-value judgment paradigm. In J. Degen, M. Franke, & N. D. Goodman (eds.), *Proceedings of the Formal & Experimental Pragmatics Workshop*, 1-6. Tübingen.

Benz, Anton & Nicole Gotzner. In Press. Embedded implicature: What can be left unsaid? *Linguistics & Philosophy.*

Bleotu, Adina C. 2019. What colouring can tell us about the acquisition of scalar items in Child Romanian. OSF. September 29. osf.io/bwrvt.

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

Clifton Jr, Charles & Chad Dube. 2010. Embedded implicatures observed: A comment on Geurts and Pouscoulous (2009). *Semantics and Pragmatics* 3(7). 1–13. http://dx.doi.org/10.3765/sp.3.7.

Crain, Stephen & Cecile McKee. 1985. Acquisition of structural restrictions on anaphora. *Proceedings of North Eastern Linguistic Society (NELS)* 15. 94-110.

Crain, Stephen & Rosalind Thornton. 1998. *Investigations in Universal Grammar: A guide to experiments on the acquisition of syntax and semantics*. Cambridge, MA: MIT Press

Degen, Judith & Michael K. Tanenhaus. 2011. Making inferences: the case of scalar implicature processing. In L. Carlson, C. Hölscher & T. Shipley (eds.), In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. 3299–3304.

Foppolo, Francesca, Maria Teresa Guasti, & Gennaro Chierchia. 2012. Scalar implicatures in child language: Give children a chance. *Language learning and development* 8. 365-394. https://doi.org/10.1080/15475441.2011.626386.

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

Gotzner, Nicole & Anton Benz. 2018. The Best Response Paradigm: A new approach to test implicatures of complex sentences. *Frontiers in Communication* 2(21). 1-13. https://doi.org/10.3389/fcomm.2017.00021.

Heizmann, Tanja. 2006. Acquisition of deontic and epistemic readings of *must* and *müssen*. In Tanja Heizmann (ed.), *University of Massachusetts Occasional Papers in Linguistics (UMOP) 34: Current issues in language acquisition*. Amherst, MA: GLSA, UMass Amherst.

Hirst, William & Joyce Weil. 1982. Acquisition of epistemic and deontic meaning of modals. *Journal of Child Language*, 9(3). 659–666. https://doi.org/10.1017/S0305000900004967.

Katsos, Napoleon & Dorothy Bishop.2011. Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition* 120 (1). 67-81. https://doi.org/10.1016/j.cognition.2011.02.015.

Noveck, Ira. 2001.When children are more logical than adults. *Cognition* 78(2). 165-188. https://doi.org/10.1016/S0010-0277(00)00114-1.

Noveck, Ira A., Simin Ho & Maria Sera. 1996. Children's understanding of epistemic modals. *Journal of Child Language* 23 (3): 621-643. https://doi.org/10.1017/S0305000900008977.

Ozturk, Ozge & Anna Papafragou. 2015. The acquisition of epistemic modality: From semantic meaning to pragmatic interpretation. *Language Learning and Development* 11 (3). 191-214. https://doi.org/10.1080/15475441.2014.905169.

Papafragou, Anna & Julien Musolino. 2003. Scalar implicatures: Experiments at the semantics-pragmatics interface, *Cognition* 86(3). 253-282. https://doi.org/10.1016/S0010-0277(02)00179-8.

Rosch, Eleanor. 1973. Natural categories. *Cognitive Psychology* 4(3). 328–50. https://doi.org/10.1016/0010-0285(73)90017-0.

Rosch, Eleanor. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology* 104(3):192–233. https://doi.org/10.1037/0096-3445.104.3.192.

van Tiel, Bob. 2013. Embedded scalars and typicality. *Journal of Semantics* 31(2).147-177. https://doi.org/10.1093/jos/fft002.

Zehr, Jeremy & Florian Schwarz. 2018. PennController for Internet Based Experiments (IBEX). https://doi.org/10.17605/OSF.IO/MD832.