Izabela Skoczeń* and Aleksander Smywiński-Pohl

Numeral terms and the predictive potential of Bayesian updating

<https://doi.org/10.1515/ip-2021-2015>

Abstract: In the experiment described in the paper Noah Goodman & Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science* 5(1). 173–184, empirical support was provided for the predictive power of the Rational Speech Act (RSA) model concerning the interpretation of utterances employing numerals in uncertainty contexts. The RSA predicts a Bayesian interdependence between beliefs about the probability distribution of the occurrence of an event prior to receiving information and the updated probability distribution after receiving information. In this paper we analyze whether the RSA is a descriptive or a normative model. We present the results of two analogous experiments carried out in Polish. The first experiment does not replicate the original empirical results. We find that this is due to different answers on the prior probability distribution. However, the model predicts the different results on the basis of different collected priors: Bayesian updating predicts human reasoning. By contrast, the second experiment, where the answers on the prior probability distribution are as predicted, is a replication of the original results. In light of these results we conclude that the RSA is a robust, descriptive model, however, the experimental assumptions pertaining to the experimental setting adopted by Goodman and Stuhlmüller are normative.

Keywords: machine learning; numerals; rational speech act model; scalar implicature

1 Introduction

1.1 Scalar implicatures

The notion of implicature was introduced early by Paul Grice (1975), who noticed that through uttering sentences in context, people convey more than just the literal meaning. Consider the following example:

*Corresponding author: **Izabela Skoczeń**, Faculty of Law and Administration, Jagiellonian University and Jagiellonian Centre for Law, Language and Philosophy, Krakow, Poland, E-mail: izabela.skoczen@uj.edu.pl

Aleksander Smywiński-Pohl, AGH University of Science and Technology, Krakow, Poland

A: Are you hungry?

B: I have had breakfast.

Imagine that this conversation takes place before noon. Although, literally taken, B's reply doesn't say whether B is hungry or not, the conversation is perfectly understandable. This is because it is possible to infer from B's reply that she is not hungry. Thus, the implicature carried by B's answer will be that "B is not hungry." Moreover, the speaker can cancel implicatures explicitly. Consider the following example:

A: Are you hungry?

B: I have had breakfast, but I am still hungry.

The second part of B's reply is an explicit cancellation of the implicature formed on the basis of the first part of the sentence, namely that "B is not hungry."

Grice distinguished between two types of conversational implicatures. First, particularized conversational implicatures (PCIs) are very strongly context dependent, just as in the example above. Second, generalized conversational implicatures (GCIs) arise in most contexts since they rely on the words used. Scalar implicatures discussed in the present paper are standard examples of the GCIs discussed in the literature (Grice 1989).

Numeral terms examined in the present paper, such as "two", are examples of scalar terms, though it is controversial whether they can be labeled implicatures. Lexically speaking, "two" could mean "exactly two," "at least two" and "not more than two." When the speaker has full knowledge and the hearer knows that the speaker has full knowledge of the context, then by using "two" the speaker will usually be taken as conveying "exactly two". This is the "knowledge inference" for numerals, we will call it the "upper bound inference for numeral term". It is based on the assumption that if the speaker had meant more or less than "two," she would have used a word that is higher on a lexical scale (for instance "one") or lower on the scale (for instance "three") (Horn 2006). Thus, the scale, or a "totally ordered set of lexical items which vary along a single dimension" (Bergen et al. 2016), is this:

(1.) <one, two, three, four, five, etc.>

By contrast, the inferred exactness should not occur when the speaker does not have full contextual knowledge (Goodman and Stuhlmüller 2013). Imagine a party to which three guests have been invited; the speaker knows that two of the guests

have left the party, yet the speaker does not know whether the third guest has left the party. The speaker says:

(II.) Two of the guests have left the party.

If the hearer is aware of the speaker's partial knowledge, the hearer should not infer "exactly two." Rather, the hearer should form the uncertainty inference¹ "at least two of the guests have left."

There is a heated debate on the nature of scalar inferences. This debate is due to an unclear border between the semantics and pragmatics of natural language as well as the unclear role of epistemic reasoning in deriving scalar inferences. Three main camps can be distinguished (Sauerland 2012).

First, the lexical camp claims that scalars are stored in the lexicon and retrieved in a relevant context. This camp can be further subdivided into three variants. The first main claim dominating in the literature is that the lexical entry consists of the "exactly" meaning of numerals (for example "exactly two"), while the "at least" or "at most" readings are the result of pragmatic mechanisms (for instance scalar implicature) cf. (Breheny 2008).

The second claim is that pragmatic enrichment happens already at the lexical level (Chierchia 2004; Levinson 2000). The third claim is that the lexical entry is underspecified and requires precisification at the pragmatic level (Carston 1998).

Second, the grammatical camp claims that there is a covert exhaustification operator "only" that triggers scalar inferences, without resort to reasoning about the epistemic state of the speaker (Chierchia 2006; Chierchia et al. 2011; Fox 2007). The asset of this camp is that it explains neatly the ambiguity of numerals (Spector 2013).

¹ The debate on numeral terms differs from the debate on quantifier terms such as 'some' and recent experimental work points to the direction that quantifiers and numerals might need distinct explanations. In the Gricean framework, the exact reading of 'some' is a scalar implicature, and under the pragmatic approach, this implicature is derived if the speaker's competence can be assumed. In the case of 'some', one can identify mentions of primary and secondary implicature, where the primary implicature is weaker and uses the belief operator directly acting at the stronger alternative and under the scope of negation: not B_S (all), the secondary implicature assumes speaker competence: B_s \not (all). If the speaker has partial knowledge, the hearer does not infer the scalar implicature (or infers the primary implicature). If no implicature is derived for numerals, they are interpreted under the at least reading: "at least two of the guests left" which is the literal, semantic, interpretation (under the Gricean analysis). This is different however, in the "exactly semantics" view, where the "at least some" is pragmatically derived. To sum up, the formulations "knowledge and uncertainty inference" are used here only for ease of exposition of the study design as they depend on what general assumptions one has about the nature of "some" and numeral term inferences (cf. Dieuleveut et al. 2019).

Third, there is the pragmatic camp. Those proponents of the pragmatic camp, which operate within the Gricean framework, rely on the competence assumption of the speaker – the speaker wants to be as informative as possible given the knowledge of the context that she disposes of. The standard account based on lexical scales is part of the pragmatic camp (Atlas and Levinson 1981; Geurts 2010; Horn 1972; Russell 2006; Sauerland 2004; Spector 2003; and others). Relevance theorists are also part of the pragmatic camp, though they do not directly employ the notion of scales (cf. Sperber and Wilson 2006). Finally, the Rational Speech Act model (RSA) discussed in the present paper is equally part of the pragmatic camp.

1.2 The RSA model

The present study is an attempt to replicate N. Goodman and A. Stuhlmüller's experiment that provided empirical support for the predictive power of the Rational Speech Act (RSA) model (Goodman and Stuhlmüller 2013). This is a model based on game theory and Bayesian decision theory. The model aims at predicting the hearer's utterance interpretation. This prediction is based on whether one updates her beliefs about the probability distribution of the occurrence of an event with the received information (Bergen et al. 2016; Frank and Goodman 2012; Goodman and Frank 2016). Interestingly, it remains unclear to what extent the model differentiates predictions of beliefs about what has been said as well as beliefs about the state of the world. Testing this requires devising an experiment where the listener does not fully trust the speaker and beliefs about what has been said as well as the state of the world diverge (Skoczeń and Smywiński-Pohl, forthcoming).

The experiment replicated in the present paper was supposed to provide empirical support for a version of the RSA model, which models belief updating (or lack thereof) in contexts where both the speaker and hearer do not have full knowledge of the circumstances relevant for the utterance. The RSA is composed of a model of the listener, a model of the speaker as well as a utility function. Let us now discuss these elements as found in the (Goodman and Stuhlmüller 2013) paper.

The listener is modeled as performing reasoning about the state of the world s , given the information she disposes of, namely, the utterance w she has heard and given the “access of the speaker” a . The “access of the speaker” is the number of objects (here out of three), which the speaker checked (so that she knows whether these objects have the relevant property). This ratio is proportional to the speaker's

model (P_{speaker}) factored by the listener's prior probability distribution on states of the world – $P(s)$. These are the beliefs of the listener's about how probable a state of the world is before hearing the speaker's utterance:

$$P_{\text{listener}}(s|w, a) \propto P_{\text{speaker}}(w|s, a)P(s)$$

The speaker is modeled as performing reasoning about which utterance to choose given the observation o she made as well as the access a she had (the number of objects (here out of three), which the speaker checked). This ratio is proportional to the product of the α parameter, the utility function U as well as the expectation ranging over the speaker's belief states $E_{P(s|o, a)}$:

$$P_{\text{speaker}}(w|o, a) \propto \exp(\alpha E_{P(s|o, a)}[U(w; s)])$$

The alpha parameter (also labeled in machine learning the soft-max optimization function or the Luce choice rule) permits the speaker to choose the utterance that will transmit the most information in the entire course of communication (not only in a one-shot language game). The expectation ranges over the beliefs of the speaker, because the speaker is not certain in which world she actually is (in the case she does not have full access to the state of the world). The beliefs of the speaker consist of a probability distribution over possible states of the world given the observation the speaker made (how many objects have the relevant property) and the access. The utility function models what information the utterance conveys in a world state:

$$U(w; s) = \ln(P_{\text{lex}}(s|w))$$

The utility function is defined as a probability distribution over states of the world given the literal meaning of the utterance, “related to the amount of information that a literal listener would not yet know about state s after hearing it described by utterance w ” (Goodman and Stuhlmüller 2013). Thus, this probability distribution is determined by the literal meaning, which is assumed to be a set of truth functions for each utterance. Since the listener does know the speaker's access, but does not know what observation the speaker made, the speaker's model becomes:

$$P_{\text{speaker}}(w|s, a) = \sum_o P_{\text{speaker}}(w|o, a)P(o|a, s)$$

Note that the costs of utterance production have been neglected here. Let us now turn to the question of probability distributions prior to receiving information on an event as described in the RSA model.

1.3 The RSA model – descriptive or normative?

Aristotelian logic could be viewed as one of the first attempts to model human reasoning. Since the deductive reasoning requirements and the assumption of monotonicity proved to generate conclusions too far stretched from the conclusions formulated in everyday reasoning, the debate turned to non-monotonic logics as well as probabilistic accounts of reasoning. In other words, the debate turned to the idea that humans think in terms of probability distributions of events happening, since a good number of every-day contexts are uncertainty contexts. Linguistic utterances provide information which allows to update probability distributions of how likely it is that an event will occur. The formal basis for such updating is labeled the Bayes Theorem (cf. for an overview Oaksford and Chater 2009). The question which arises again is whether, analogously to the fall of the model of deductive reasoning as mirroring human inference, the Bayesian model is a good model of human linguistic inference.

Roughly, the spectacular fall of deductive models was based on the fact that the conclusions arising out of such reasonings were different than the conclusions arising out of everyday human reasoning. If the Bayesian approach is to be taken as the descriptively adequate approach, then the minimal requirement is that it generates conclusions conform to conclusions reached by humans. Otherwise, the model becomes a normative model which could be at best described as wishful thinking rather than a description of the human cognitive processes. This of course depends on the aim of the model. For instance, the currently blooming field of conceptual engineering aims at such normative, ideal, rather than descriptive models (Koch 2020). Yet this is clearly not the case with the Rational Speech Act model, which, as its proponents claim, is supposed to *explain* how, in practice, human language inference is heavily reliant on context:

(...) in practice, the meaning we derive from language is heavily dependent on nearly all aspects of context, both linguistic and situational. To formally *explain* (emphasis added) these nuanced aspects of meaning and better understand the compositional mechanism that delivers them, recent work in formal pragmatics recognizes semantics not as one of the final steps in meaning calculation, but rather as one of the first. Within the Bayesian Rational Speech Act framework (Frank and Goodman 2012), speakers and listeners reason about each other's reasoning about the literal interpretation of utterances (Scontras et al. 2018).

If the RSA is a descriptive rather than normative model, then a minimal prerequisite to uphold the model as a descriptive model of human language processing is that the model yields the same output as the output of human reasoning in the contexts the RSA models. As the proponents of the RSA advance, the model's

predictions about the output of reasoning are empirically testable. One can simply ask participants what they would understand from an utterance in an uncertainty context. This is especially interesting as the putative success of this enterprise not only strengthens the RSA model but also the more general claim that human reasoning in uncertainty contexts is Bayesian (or at least can be predicted by Bayesian mechanisms, cf. Baron 2006).

Crucially, while the literature on RSA provides plenty of experiments confirming the model's predictions with survey data (Bergen et al. 2016; Frank and Goodman 2012; Goodman and Frank 2016; Yoon et al. 2020), there is also an extensive literature on the market which criticizes the Bayesian approach to modeling human reasoning as descriptively inaccurate (cf. for instance Baratgin 2009; Baratgin and Politzer 2006; Evans et al. 2002; Krynski and Tenenbaum 2003; Stanovich and West 1998). In light of this debate, it is vital to replicate (also cross-culturally) the findings of the RSA model, as it can shed light on two main issues. First, whether the RSA's predictions are robust in the sense they generally arise in the contexts which the RSA describes. Second, whether human linguistic inference in uncertainty contexts is indeed mirrored by Bayesian mechanisms. Since the Bayes Theorem is a mathematical theorem, numeral terms seem a natural first candidate to inquire about.

The original experiment by Goodman and Stuhlmüller was devised in two versions. The first version employed the quantifier term “some”. We have replicated this experiment and discussed the divergent results in a separate paper, where we posed a different hypothesis concerning ambiguity limited to quantifier terms (Skoczeń and Smywiński-Pohl, ms). The second version of the original experiment employed numeral terms and this version is replicated in Polish in the present paper.

As stated above, as a Bayesian model, the RSA predicts an interdependence between beliefs about the probability distribution of the occurrence of an event prior to receiving information and the updated probability distribution after receiving information. This is reflected by the structure of the listener's model:

$$P_{\text{listener}}(s|w, a) \propto P_{\text{speaker}}(w|s, a)P(s)$$

The listener reasons about the state of the world s , given the utterance w she has heard and the access a that captures the amount of information she has about the context of utterance (Goodman and Stuhlmüller 2013). This is proportional to the speaker's model (who reasons about which utterance w she should choose given the state of the world s and the access to context a) that is, crucially, factored by the beliefs about the prior probability distribution of the possible state of the world $P(s)$. This dependence on priors occurs also within

the models of higher-order listeners, that is listeners after several iterations of the model.²

We decided to test whether this interdependence would be cross linguistic and would occur in a different language, namely Polish. Consequently, the general aim of the paper is to check whether the interdependence predicted by the RSA model between the prior and posterior probability distribution of objects having a property is psychologically adequate. In other words, we are checking whether it is the case that only when participants' answers on the prior probability distribution are as predicted, then (and only then) the answers on the posterior probability distributions are also as predicted. If this is not the case, we check what factors prevent it, investigating whether Bayesian updating is indeed an adequate psychological model of human numeral term processing.

Moreover, we have two specific aims based on two main elements that raise our concerns in the empirical project of testing the RSA's predictions with experiments. The first element is the experimental design and results analysis employed by Goodman and Stuhlmüller. The second element is the construction of the model itself, namely, the factoring of results with binominal distribution parameters.

Moving to the first element, we fear that the complex experimental design will not be clear to most participants. Most certainly, it is an established experimental practice to exclude participants who are not attentive enough, do not pass a simple comprehension question on the scenario, take too little time to answer questions (which hints that they did not even read the questions) or are not native speakers of the language employed in the survey cf. (Barnett 1994). However, such practice is licensed only if it aims at removing outliers. By contrast, if too many data points are removed, the results of the experiment can describe the reasoning of merely a skewed part of the population rather than a general representative sample. If the RSA is to be a descriptive model, then the majority of participants should answer as the RSA predicts. If this is not the case, then the RSA might either model a sub-part of the population or be treated as a normative model. Investigating which sub-part of the population reasons as the RSA predicts can also provide a window into the mechanisms of specific human cognitive processes.

Another concern related to the experimental design, pertains to the computation of the parameters of the model. The RSA model employed in the explanation

² Recent empirical research suggested that the interdependence of priors and posteriors predicted by the RSA does not always find support in people's reactions, namely, it does not find support in cases where manipulated priors are extreme, e.g., an utterance of "some" when "all" is extremely likely *a priori* (e.g., "Some of the triangles have three sides") (cf. Degen et al. 2015).

of the reasoning of the experiment's participants has two parameters: the base probability rate (p) and the speaker optimality parameter (α). The first parameter is used to compute the prior distribution of events, i.e. it is a parameter of the binomial distribution, used to calculate the literal meaning of the utterance. The second parameter is used in the softmax function to allow the speaker to select the appropriate amount of information to convey. These parameters are determined in the original experiment by optimizing the Root Mean Square Error (RMSE) between **the means of the model's predictions and the means of the participants posterior decisions**, i.e. the procedure directly reduces the difference between them.

Such an approach is very common in Machine Learning (ML), when the model's parameters are optimized to best predict the (training) dataset. Yet in ML the computed parameters are later applied to predict the distribution of an unseen sample of data (the test set). Reporting the error rate (here RMSE) on the training set in ML could be viewed as suspicious. Using the optimization procedure to estimate RSA model parameters, seems valid, if we assume that the model is appropriate and we want to explain some feature of the model (e.g. we want to compare the prior distribution of events between the model and the participants or between two groups of participants). Yet if the optimized values are used to judge that the model is valid, since the predictions of the model **with the optimized parameters** are similar to the observed means, the procedure seems to be controversial.

Let us now proceed to the replication of Goodman and Stuhlmüller's experiment employing numeral terms.

2 Experiment 1 – replication in the Polish language with varied age groups

2.1 Participants

Subjects were recruited through a Polish online survey platform “Research online.” Just as in the original experiment performed by Goodman and Stuhlmüller, we excluded participants who were not native speakers of Polish, who responded incorrectly to the training questions as well as a comprehension control question. Research online guaranteed that no participant took the study twice. 62% of participants were female, the age range of the sample was 18–45+ years, while the mean age was 41 years.

2.2 Methods and materials

Just as in the original experiment, we used the same 6 scenarios: letters with checks inside, students passing an exam, fruits with dried pith, mobile phones with broken transistors, sprouting seeds and winning lottery tickets (for exact scenarios cf. Appendix). The study began with two warm-up control questions checking the general attention of the participants and familiarizing them with a betting measure. Participants had to divide a hundred money units by betting on zero, one, two or three objects having the property. The scenarios were displayed in a randomized order. Each scenario was about three objects and a potential property that these objects could have, for example:

Letters to Laura's company almost always have checks inside. Today Laura received 3 letters.

Independently of scenario, participants always saw the information that the objects “**almost always**” have the property in question. The aim was to increase participants' belief that it is likely that all objects have the property. This way, after participants were presented with the utterance, we could distinguish an upper bound for a numeral term from a prior belief that it is unlikely that all objects have the property.

Next, we asked a question to measure participants' beliefs on the probability distribution of objects having the property before hearing the utterance. We call this the prior probability distribution:

How many of the 3 letters do you think have checks inside?” (Goodman and Stuhlmüller 2013)

After capturing the prior probability distribution on the objects having the property, we displayed the information about the number of objects verified by the speaker plus the number of objects, out of the ones that the speaker checked, that have the property. The number of verified objects with the property in question varied from one to three:

Laura tells you on the phone: “I have looked at 2 of the 3 letters. Two of the letters have checks inside”. (Goodman and Stuhlmüller 2013)

Next, a question about the posterior probability distribution and a control question followed.

Now how many of the 3 letters do you think have checks inside?

Do you think Laura knows exactly how many of the 3 letters have checks inside? (Goodman and Stuhlmüller 2013)

Since the speaker's access varied, she verified, for instance, one, two or three letters; each participant was presented with each access condition in a random order with randomly chosen scenarios (participants never saw the same scenario twice). Just as in the original experiment, there were three partial-knowledge conditions (without knowledge inference) and three complete-knowledge "control" conditions (with knowledge inference).

Using standard terminology, this is a within subjects design (all participants see all the 6 scenarios in a randomized order). Moreover, every scenario contains a different condition, so each participant sees all of the six conditions. This is because, within each displayed scenario we randomize the number of objects seen by the speaker (what is labeled access): the speaker may have verified one, two or three objects out of three. Finally, we randomize how many of the verified objects have a property, which is conveyed through the speaker's utterance in each scenario:

- A. If the speaker verified one object out of three, she could only say that "one of the objects has the property in question".
- B. If the speaker verified two objects out of three, she could say either that (a) "one of the objects has the property" or (b) "two of the objects have the property" (we randomized the display of these two utterances).
- C. If the speaker verified three objects out of three, she could say either that (a) "one of the objects has the property" or (b) "two of the objects have the property" or (c) "three of the objects have the property" (we randomized the display of these three utterances).

The 3 possible utterances in A and B are the partial knowledge conditions, while the 3 possible utterances in C are the complete knowledge conditions (because the speaker checked all three objects out of three objects so she has information concerning each and every object: whether it has the relevant property).

Since there is a control question within each displayed condition to a participant, just as in the original experiment, we consider each condition displayed to the participant as a separate trial. Thus, for each participant we can exclude the data from conditions where the participant answered the control question incorrectly and employ only data from conditions where the participant answered the control question correctly (just as in the original experiment, we think however this is a controversial procedure and elaborate on the matter in the next sections). Thus, the statistics presented in the present paper are per trial statistics.

The experimental design is presented in Figure 1 below.

Figure 1: The experimental design. The order of appearance was randomized with an even presentation of elements.

2.3 Results

A repeated measure ANOVA, where the dependent measure was the bets on 3 and the independent measure was scenario (6 possibilities), showed no significant effect ($F(5,294) = 1.06$, $p = 0.382$, $\eta^2 = 0.02$). Thus, different scenarios did not affect the answers, so we will not include scenario effects in the analyses.

Contrary to predictions, the answers on the control question in each condition were not always in accordance with the statement of the speaker concerning the number of objects she had verified. When the speaker declared she verified all of the objects, there were participants who answered the question on whether the speaker knows exactly how many objects have the property in the negative. Conversely, when the speaker declared she did not check all of the objects, there

were participants who answered positively the question on whether the speaker knows exactly how many objects have the property in question: bets that speaker had complete knowledge in partial-access conditions, $M = 46.9$, $SD = 42.9$; in complete-access conditions, $M = 86.0$, $SD = 28.2$ (compared to the original, $M = 42$ $SD = 3.4$ and $M = 92.1$ $SD = 1.6$). Just as in the original experiment, each participant saw all 6 conditions. Moreover, just as in the original study, we excluded answers from conditions in which the participants in their answers on the control question (for instance: “Do you think Laura knows exactly how many letters have checks inside?”) bet less than 70 on the “yes” answers in the complete access conditions (=the speaker checked all three objects) or bet less than 70 on the “no” answer in the incomplete access conditions (=the speaker did not check all the three objects).

We perform these exclusions in line with the results analysis of the original experiment. The threshold 70 could be considered arbitrary. Goodman and Stuhlmüller argue that the purpose of this exclusion of participants is to make sure that we analyze the data of participants who understood which precise situation is being investigated. If the protagonist claims that he saw 3 out of 3 objects, the experimental goal is that participants would take the protagonist as knowing how many objects have the property. With a betting measure, a bet over 70 is a clear indication that the answer is positive, a counterpart of being significantly above the mid-point 4 in a 7 point Likert scale. This way there is ground to claim that participants did not think that the protagonist either did not check the objects carefully enough to know whether they have the property, or, is lying that he indeed checked all the objects. Thus, according to Goodman & Stuhlmüller it is not an arbitrary exclusion, but one that serves the purpose of guaranteeing that participants understood the scenario and were attentive. In order to ascertain that this is the case, we recalculated all experimental data without filtering participants with the control question on knowledge. We obtained the same results in terms of significance tests: see Appendix.

For the analysis below, if the protagonist said that he did not see all the objects, we therefore, on the same grounds, excluded those who bet less than 70 on the claim that the protagonist does not know whether all the objects have the property in question. We acknowledge however that in such situations, in real life rather than laboratory ones, the speaker might be less trustworthy, and thus the assessment of his knowledge might be different than the one assumed by the experimental purpose.

Figure 2 presents mean participant bet on each world state, varying the word the speaker used and the speaker’s perceptual access. The formulation “1 access 1” means that the speaker has access to information concerning the properties of one object out of three and utters “one.” “Access 2” means the speaker has access to information concerning the properties of two objects out of three; “access 3” means

Figure 2: Replication with the participation of 50 subjects from varied age groups: mean participant bet on each world state (prior on the left, posterior on the right), varying the word the speaker used and the speaker's perceptual access. Data have been filtered to include only trials in which the participant's bet that the speaker had complete knowledge was greater than 70 in the expected direction. Error bars represent the standard error of the mean.

the speaker has access to information concerning the properties of three objects out of three. In the “one access three” condition we excluded 9 trials (leaving 41), in the “two access three” we rejected 11 trials, in the “three access three” condition we rejected 6 trials. In the “one access one” condition we rejected 27 trials, in the “one access two” 27 trials and in the two access two 31 trials.

To evaluate the overall effect of access, we performed an ANOVA with access and word as independent measures and bets on 3 as the dependent measure. Each trial has a different number of participants (since we filtered out participants with a control question for each trial separately) and thus each pairwise t-test comparison has different degrees of freedom.

We found an effect of access ($F(2,18) = 3.33$, $p = 0.038$, $\eta^2 = 0.035$ (in the original study $p < 0.001$)), however, in contrast to the original study, we found no interaction between word and access ($F(1,18) = 1.13$, $p = 0.290$, $\eta^2 = 0.006$ (in the original study $p = 0.006$)). However, we found an effect of word ($F(2,18) = 101.55$, $p < 0.001$, $\eta^2 = 0.526$).

We then explored the results in more detail using planned comparisons to test whether inferences of upper bounds for numeral terms were drawn (only) when predicted. Before proceeding, one important clarification: in the RSA, upper bound inference (misleadingly labeled implicature inference in the original study), is not binary. In other words, it is not the case that one either draws or refrains from the inference at stake. Rather, pragmatic inferences are a matter of degree and probability distributions. Consequently, the terminology (implicature versus partial

implicature) employed in Goodman & Stuhlmüller's study is not fully accurate. We should rather speak in terms of inference strength or probability distribution. Moreover, as discussed above, it is controversial whether in the case of numerals we can speak of implicatures at all. Thus, we employ the term "upper bound inference for numerals". Moreover, we will discuss the degrees of strength of such inferences rather than treat them as binary measures.

We found a strong tendency toward upper bound inference in the complete-access conditions: when the speaker said "two," bets on state 3 were less than on state 2 (paired, directional t-test, $t(39) = 7.96$, $p < 0.001$, $d = 1.27$). In other words, participants bet less on the possibility that all objects have the property in question, than on the possibility that not all objects have the property in question.

When the speaker said "one," bets on state 1 were greater than on state 3 (paired directional t-test, $t(40) = 8.98$, $p < 0.001$, $d = 1.44$) or state 2 (paired directional t-test, $t(40) = 10.34$, $p < 0.001$). The effect size was $d = 1.62$.

In contrast, there was no tendency toward upper bound inference when access was 1 and the speaker said "one": bets on 1 were not greater than on 2 (paired directional t-test, $t(22) = 0.25$, $p = 0.403$, $d = 0.05$) or on 3 (paired directional t-test, $t(22) = 1.39$, $p = 0.090$, $d = 0.29$). In other words, participants bet similar values on all and not all objects having the property in question.

However, in contrast to the original study, there was a strong tendency toward upper bound inference when access was 2 and the speaker said "two": bets on 2 were greater than on 3 (paired, directional t-test, $t(18) = 3.37$, $d = 0.77$, $p < 0.001$ (compare to the original study, in which no tendency toward upper bound inference was found since $p=0.870$)). For this reason, we cannot consider the present results a full-fledged replication of the original study. Nevertheless, we think that a possible cause for the failure to replicate the result at stake is due to the way in which participants answer the question on the prior probability distribution. In the "Two access two condition," participants' mean bets on priors were the following: zero = 3.42, one = 19.21, two = 63.15, three = 14.21. Thus, the highest bets were on two as if already *ex ante*, before receiving information, there was a belief in low probability of all objects having the property in question.

When access was 2 and the speaker said "one," we found the predicted by RSA weaker tendency toward upper-bound inference in the probability distribution: bets on state 1 were significantly bigger than on state 3 (paired directional t-test, $t(22) = 2.42$, $p = 0.012$, $d = 0.50$), and on state 2 (paired directional t-test, $t(22) = 1.85$, $p = 0.039$, $d = 0.39$).

In order to balance the sample and avoid overrepresentation of a particular age group, we tested different age groups as part of our full sample. As a secondary analysis, in search for potential reasons for the divergent results (compared to the original paper), we checked the effect of age on this full sample. There was an effect

of age: the first age group's (18–25 years old) mean bets on two in the “two access two” condition were: $M = 85.0$, $SE = 2.9$, in the second age group (26–44 years old) $M = 91.7$, $SE = 5.4$ and in the third age group (45+): $M = 49.4$, $SE = 13.2$. Differences in bets on two, in the “some access two condition”, depending on age groups: ANOVA $F(5,13) = 3.94$, $p = 0.021$, $\eta^2 = 0.600$.

Corrections for multiple comparisons: to avoid the problem of increasing alpha error we performed a Holm's correction for multiple comparisons (Holm 1979). There were no differences in significance (Gaetano 2013). See Appendix Table 2.

As a final step we have compared the results of the experiment with the RSA model predictions by following the same procedure as in the original research. The model has two parameters: base rate (p -base) of the binomial distribution, used to predict the prior distribution of bets and speaker optimality parameter α . These parameters are estimated by minimizing the Root Mean Square Error (RMSE) between the model predictions and the bets provided by the participants. As we argued earlier, the predictions of the model based on the optimized parameters, may not be used to judge if the model is coherent with the observation. To mitigate this issue, we have decided to modify the procedure, to align it with the machine learning best practices. In our analyses we have split the dataset randomly into halves, optimized the parameters on one half and compared the predictions with the other half (called “blind” in the remaining part of the article). We ensured that bets of each participant were present only in one of the halves. The computed parameters were the following: p -base = 0.42, $\alpha = 0.62$, RMSE = 9.8 (original research: RMSE = 9.01). The RMSE between the means of the model and the blind group was 11.56. The results of the comparison between the model predictions and the bets of the blind group are given in Figure 3.

Figure 3: The comparison of the posterior bets of the blind group of participants (left) with the predictions of the RSA model (right) for the first experiment.

2.4 Discussion

Differences between the original study and experiment 1 appeared in the “two access two” condition. This difference was crucial since in the original experiment the inference “not more than two” did not appear, while in the data we collected this inference did appear. In other words, the bets on “two” and “three” in the “two access 2” condition were similar in the original experiment carried out by N. Goodman and A. Stuhlmüller, while in the replication in Polish the bets on “two” were significantly higher than the bets on “three” (see Figure 4).

However, an important factor that contributed to this result could have been the imperfect prior probability distribution manipulation. It could be that participants ignored the information that the objects almost always have the property in question and bet less on all of the objects having this property. This observation is supported by the p -base parameter estimated for the model using the optimization procedure, The value 0.42 is below 0.5, which hardly can be interpreted as “almost always”.

In an attempt to mitigate this issue, we modified in experiment 2 the way information was displayed. Namely, after the information on prior probability was displayed (that the objects almost always had the property in question), the participants had to press a button labeled “next” to be able to display the first question. This way we hoped to make more salient the information that the objects “almost always” had the property in question.

Figure 4: Differences between the original experiment (to the left, reuse license number 4343821165053, source: <https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12007>) and experiment 1.

Nevertheless, it could also be that the “almost always” information was interpreted as “not always.” In other words, “almost always” was not interpreted as predicted; namely, it was not interpreted as “there is a high probability that all of the objects have the property,” but rather, as “there is a high probability that two out of the three objects have the relevant property.” Thus, we suspect that the intended prior manipulation is only partly successful because the phrase:

(I.) Letters to Laura’s company almost always have checks inside.

is enriched to the interpretation “not always.” This result is puzzling in the sense that the Polish translation of “almost always” has no semantic differences with the English counterpart. Perhaps Goodman and Stuhlmüller’s assumption on the interpretation of the “almost always” phrase was normative rather than descriptive.

We suspected that the experimental design was complex for participants and put quite high demands on their attention, which could be better achieved by students. Some of the participants commented that the questions and scenarios were “almost the same.” Thus, it could be the case that they entered the same results, even though the scenarios they evaluated changed.

Comparison of the model predictions with the bets provided by the blind group for the original experiment and our replication may not be pursued directly, since we have changed the procedure. Still, it is apparent that the model predictions in our first experiment are not well aligned with the blind group: RMSE for the first group is 9.8, while for the blind group it is 11.56. Visual inspection of the plots given on Figure 3, also indicates, that for “two access two” and “one access two” conditions the differences are huge. Yet the model was able to capture the primary difference between the original experiment and the replication, namely the higher bet on 2 in the “two access two” condition, although the difference estimated by the model was much smaller. On the other hand, the ranking of bets in conditions “one access one” and “one access two” is different for the model and for the blind group. As a result, we think that the judgment that the RSA model captures the reasoning of the participants is not supported by this experiment.

Before proceeding, a final remark: the results of experiment 1 depict that a large number of trials had to be discarded and a good number of manipulations was not fully successful. This opens the question whether the predictions generated by the RSA model are indeed results that mirror human reasoning, or, rather, they are some laboratory idealizations.

3 Experiment 2 – replication in the Polish language with students' participation

3.1 Participants

Participants were recruited through a Polish online platform “Research online” and performed the experiment for a small payment. Fifty persons participated all of whom were students currently enrolled at the Jagiellonian University. Just as in the original experiment performed by Goodman and Stuhlmüller, each study began with two warm-up control questions checking the general attention of the participants and familiarizing them with a betting measure. We excluded participants who were not native speakers of Polish, who failed the attention check, responded incorrectly to the control comprehension question or took the study twice. Seventy percent of participants were female, mean age was 32 years.

3.2 Methods and materials

We employed the exact same methods and materials as in the original experiment and in experiment 1 with one exception. In an attempt to mitigate the incorrect prior manipulation issue, we modified the way information was displayed. Namely, after the information on prior probability was displayed, participants had to press a button labeled “next” to be able to display the first question.

The experiment was performed in the Polish language. Since we hypothesized that the experimental design was complex for participants, we decided to test only students and check whether higher levels of attention deployed by students would influence the results.

3.3 Results

A repeated measure ANOVA, in which the dependent measure were bets on 3 and the independent measure were scenarios (6 possibilities), showed no significant effect of scenario ($F(5,17) = 0.92$, $p = 0.472$, $\eta^2 = 0.030$). Thus, different scenarios did not affect the answers, so we will not include scenario effects in the analyses.

Contrary to predictions, the answers on the control question in each condition were not always in accordance with the statement of the speaker concerning the number of objects she had verified. When the speaker declared she verified all of the objects, some participants answered the question on whether the speaker

knows exactly how many objects have the property in the negative. Conversely, when the speaker declared she did not check all of the objects, some participants answered positively the question on whether the speaker knows exactly how many objects have the property in question: bets that speaker had complete knowledge in partial-access conditions, $M = 28.2$, $SD = 37.4$; in complete-access conditions, $M = 88.7$, $SD = 29.0$ (compared to the original, $M = 42.0$ $SD = 3.4$ and $M = 92.1$ $SD = 1.6$). Just as in the original experiment, each participant saw 6 conditions. We excluded answers from conditions in which participants, in their answers on the control question (for instance: “Do you think Laura knows exactly how many letters have checks inside?”), bet less than 70 on the “yes” answers in the complete access conditions (the speaker verified all three objects) or bet less than 70 on the “no” answer in the incomplete access conditions (the speaker did not check all the three objects).

Figure 5 presents mean participant bet on each world state, varying the word the speaker used and the speaker’s perceptual access. The formulation “One access 1” means that the speaker has access to information concerning the properties of one object out of three and utters “one.” “Access 2” means the speaker has access to information concerning the properties of two objects out of three; “access 3” means the speaker has access to information concerning the properties of three objects out of three. In the “one access one” condition we rejected 17 trials, in the “one access two” condition we rejected 19 trials, in the “two access two” condition we rejected 22 trials, in the “one access three” condition we rejected 8 trials, in the

Figure 5: Experiment 2: mean participant bet on each world state, varying the word the speaker used and the speaker’s perceptual access. Data have been filtered to include only trials in which the participant’s bet that the speaker had complete knowledge was greater than 70 in the expected direction. Error bars represent the standard error of the mean.

“two access three” we rejected 8 trials, in the “three access three” condition we rejected 4 trials. Thus, most of the datasets rejections were due to a failure to pass the control question. As such a high number of excluded participants seemed suspicious, we reanalyzed the entire dataset without these exclusions, but found no difference in the significance tests (cf. Appendix).

To evaluate the overall effect of access, we performed an ANOVA with access and word as independent measures and bets on 3 as the dependent measure. Each trial has a different number of participants (since we filtered out participants with a control question for each trial separately) and thus each pairwise t-test comparison has different degrees of freedom. We found a main effect of access ($F(2,22) = 138.06$, $p < 0.001$, $\eta^2 = 0.558$), an interaction between word and access ($F(1,22) = 62.86$, $p < 0.001$, $\eta^2 = 0.223$ (in the original study $p < 0.001$)), and a main effect of word ($F(2,22) = 424.66$, $p < 0.001$, $\eta^2 = 0.795$).

Just as in the original study, we next used comparisons to test whether knowledge inferences were drawn (only) when predicted. We found such inference in the complete-access conditions: when the speaker said “two,” bets on state 3 were smaller than on state 2 (paired directional t-test, $t(44) = 23.84$, $p < 0.001$). The effect size was $d = 3.56$. When the speaker said “one,” bets on state 1 were bigger than on state 3 (paired directional t-test, $t(41) = 34.99$, $p < 0.001$). The effect size was $d = 5.40$. Moreover, when the speaker said “one,” bets on state one were also bigger than on state 2 (paired directional t-test, $t(41) = 32.96$, $p < 0.001$). The effect size was $d = 5.09$.

In contrast, there was no upper-bound of the numeral term when access was 1 and the speaker said “one”; bets on 1 were not greater than on 2 (paired directional t-test, $t(32) = -1.88$, $p = 0.069$, $d = 0.33$); however, note that our results pointed to a lower significance, while the original was $p = 0.960$. Moreover, based on the effect size of our results, there was a non-negligible difference.

When access was one and the speaker said “one,” bets on 3 were also bigger than on 1 (paired directional t-test, $t(32) = -4.53$, $p < 0.001$, $d = 0.79$).

There was no upper bound of the numeral term when access was 2 and the speaker said “two”; bets on 2 were not bigger than on 3 (paired, directional t-test, $t(27) = -0.86$, $p = 0.397$, $d = 0.16$). Note that the mean bets on prior probability distribution were as predicted: on zero objects having the property = 1.71, one = 11.75, two = 34.64, three = 51.89.

When access was 2 and the speaker said “one,” we found the predicted weaker upper bound of the numeral term: bets on state 1 were significantly greater than on state 3 (paired directional t-test, $t(30) = 5.52$, $p < 0.001$, $d = 0.99$) but not on state 2, namely bets on 2 were greater than bets on one (paired directional t-test, $t(30) = 11.55$, $p < 0.001$, $d = 2.08$).

In order to balance the sample and avoid overrepresentation of a particular age group, we tested different age groups as part of our full sample. As a secondary analysis, in search for potential reasons for the divergent results in experiment 1 (compared to the original paper), we checked the effect of age on this full sample. There was no effect of age: the first age group's (18–25 years old) mean bets on two in the “two access two” condition were: $M = 49.7$, $SE = 7.7$, in the second age group (26–44 years old) $M = 36.0$, $SE = 11.5$ and in the third age group (45+): $M = 48.0$, $SE = 24.9$. Differences in bets on two, in the “some access two condition”, depending on age groups: ANOVA $F(3,24) = 1.30$, $p = 0.296$, $\eta^2 = 0.140$.

Corrections for multiple comparisons: to avoid the problem of increasing alpha error we performed a Holm's correction for multiple comparisons (Holm 1979). There were no differences in significance (Gaetano 2013). See Appendix Table 3.

As a final step we have compared the results of the experiment with the RSA model predictions by following the same procedure as described for the first experiment (i.e. we have split the participants into halves, computed the parameters on one group and compared the results with the other group). We have obtained the following results: p -base = 0.64, α = 9.47 and RMSE = 5.36 (RMSE = 9.01 in the original research). The RMSE between the means of the model and the means of the blind group was 6.10. The results of the comparison between the model predictions and the bets of the blind group are given in Figure 6.

Figure 6: The comparison of the posterior bets of the blind group of participants (left) with the predictions of the RSA model (right) for the second experiment.

3.4 Discussion

In the discussed experiment we obtained an exact replication of the original results in the answers on both the prior and posterior probability distributions. We hypothesize that this was due to higher attention levels deployed by students. There was not enough evidence to support the hypothesis about misinterpreting the “almost always” formulation. The p -base parameter was estimated as 0.64 which is much closer to the “almost always” interpretation, than in the first experiment. Moreover, correct answers on the prior probability distribution question triggered correct answers on the posterior probability distribution question just as predicted by the RSA model.

Nevertheless, as this second experiment depicted, both a modification of the manipulation as well as a sample with higher attention levels is what is required to obtain results conform to the RSA predictions. This still begs the question: is the RSA a genuine model of human reasoning, or rather an idealized model following assumptions that its creators found to be correct in their normative sense?

The comparison between the model predictions and the bets provided by the blind group favors the interpretation that for the second experiment the RSA model captures the reasoning of the participants. The measured RMSE is 5.36 (much lower than in the original experiment) and for the blind group it is 6.10. We think that the similarity of outcomes better supports the claim of the validity of the model, since we have used half of the participants to estimate the parameters of the model and compared the predictions with the second half.

4 General discussion and conclusion

The performed experiments support the hypothesis that there is a systematic and cross-linguistic interaction between utterance understanding and world knowledge shared by interlocutors. Thus, the RSA model provides an adequate account of this interaction in communication. This in turn supports the pragmatic camp in the debate on the nature of scalar inferences.

Moreover, our results point toward the direction that, at least with respect to numeral terms, the updating is conforming to the RSA predictions. Namely, in experiment 2, where the elicitation of priors in the critical condition was as predicted (this means that bets on 3 objects having the property were greater than bets on two, and bets on two were greater than bets on one), the bets on the posterior probability distribution (after seeing information) also conformed to the predictions. By contrast, in experiment 1, in which the prior elicitation was contrary to

predictions in the critical condition, the posterior elicitation also turned out to be contrary to predictions.

Moreover, we found a relation concerning the interaction between numeral terms processing as well attention. In experiment 2, students' participants, who are deemed to exhibit higher attention levels, performed the task in conformity to the RSA model predictions. For this reason, it remains an aim for future study to devise an experiment that requires a lower level of overall attention or ensures that the attention level remains high during the answering of different questions.

To sum up, our results depict that the RSA model is robust, since if participants estimations of priors (probability distribution of objects having a property before hearing the utterance) is as assumed by the model, then participants estimation of posteriors (probability distribution of objects having a property after hearing the utterance) are conform to the RSA's predictions. However, this conclusion raises many further questions.

Even if there is a Bayesian interdependence between priors and posteriors, does this automatically mean that the RSA is a descriptive model of human reasoning? In other words, is it a formalization of theories about human language processing which is supposed to predict what inferences interlocutors will draw? Or, rather, is it a normative theory of what are the "safest" inferences to make in a certain context given the scarce information (and uncertainty) one has? This question arises because, according to Goodman and Stuhlmüller, only one of the posterior bets we collected in our experiments, namely the posterior bets from experiment 2, are correct. But what does it mean that they are correct? If the majority of the population draws the posterior inference that we found in experiment 1, then one could treat this as preliminary evidence that perhaps the RSA model is normative rather than descriptive (or represents the reasoning mechanisms of only a part of the population). We stress that this is a preliminary claim which will require further investigation.

Acknowledgments: The authors would like to thank Noah Goodman and Andreas Stuhlmüller for sharing experimental materials; the authors also thank two anonymous reviewers, Laurence Horn, Leon Bergen, Gregory Scontras, Daniel Lassiter, Krzysztof Kasperek, Anna Drożdżowicz, Katarzyna Kijania-Placek, and Krzysztof Poślajko, for comments on drafts and conference discussions.

Research funding: This research was funded by the Polish National Centre for Science (Preludium Grant No. 2015/19/N/HS5/00029).

Appendix

Scenarios and repository

Link to repository: <https://osf.io/a6utz/>.

Below are the scenarios we employed from the original study translated into Polish:

Table 1: Vignettes from experiments 1 and 2.

Nr.	English version	Polish version
1	<p><i>Seeds</i></p> <p>Corendula seeds almost always sprout within a day when put into water. Two days ago, botanist Jim put three Corendula seeds into water.</p> <p>How many of the three seeds do you think have sprouted?</p> <p>Jim tells you on the phone: “I have looked at one/two/three of the three seeds. One/two/three of the seeds sprouted.”</p> <p>Now how many of the three seeds do you think have sprouted?</p> <p>Do you think Jim knows exactly how many of the three seeds have sprouted?</p>	<p><i>Nasiona</i></p> <p>Nasiona słonecznika prawie zawsze kiełkują w ciągu jednego dnia, gdy umieści się je w wodzie.</p> <p>Dwa dni temu botanik Jan umieścił trzy nasiona słonecznika w wodzie.</p> <p>Jak myślisz, ile ziaren słonecznika spośród trzech wykiełkowało?</p> <p>Jan informuje Cię przez telefon: “Sprawdziłem jedno/dwa/trzy spośród trzech nasion.</p> <p>Jedno/dwa/trzy nasiona wykiełkowały.”</p> <p>Jak myślisz, ile spośród trzech nasion wykiełkowało?</p> <p>Czy uważasz, że Jan wie ile dokładnie spośród trzech nasion wykiełkowało?</p>
2	<p><i>Tickets</i></p> <p>Tickets in the DoubleDay instant lottery almost always win. Joe bought three DoubleDay tickets yesterday.</p> <p>How many of the three tickets do you think have won?</p> <p>Joe tells you on the phone: “I have looked at one/two/three of the three tickets. One/two/three of the tickets won.”</p> <p>Now how many of the three tickets do you think have won?</p> <p>Do you think Joe knows exactly how many of the three tickets have won?</p>	<p><i>Losy</i></p> <p>Losy na loterii fantowej prawie zawsze wygrywają.</p> <p>Janusz zakupił trzy losy na loterii fantowej wczoraj.</p> <p>Jak myślisz, ile spośród tych trzech losów wygrało?</p> <p>Janusz informuje Cię przez telefon: “Sprawdziłem jeden/dwa/trzy spośród trzech losów.</p> <p>Jeden/dwa/trzy spośród trzech losów wygrały.”</p> <p>Jak myślisz ile spośród trzech losów wygrało?</p> <p>Czy uważasz, że Janusz wie ile dokładnie spośród trzech losów wygrało?</p>
3	<p><i>Exams</i></p> <p>Students in the introductory bio class almost always have passing grade on the exam.</p>	<p><i>Egzaminy</i></p> <p>Uczestnicy kursu ‘Wprowadzenie do biologii’ prawie zawsze zdają egzamin.</p>

Table 1: (continued)

Nr.	English version	Polish version
	<p>Mark's three intro bio students took an exam yesterday.</p> <p>How many of the three exams do you think have passing grade?</p> <p>Mark tells you on the phone: "I have looked at one/two/three of the three exams. One/two/three of the exams have passing grade."</p> <p>Now how many of the three exams do you think have passing grade?</p> <p>Do you think Mark knows exactly how many of the three exams have passing grade?</p>	<p>Trzech studentów Marka podeszło wczoraj do egzaminu z 'Wprowadzenia do biologii.'</p> <p>Jak myślisz, ilu spośród tych trzech studentów zdało egzamin?</p> <p>Marek informuje Cię przez telefon: "Sprawdziłem jeden/dwa/trzy spośród trzech egzaminów."</p> <p>Jeden/dwóch/trzech studentów zdało egzamin."</p> <p>Jak myślisz, ilu spośród trzech studentów zdało?</p> <p>Czy uważasz, że Marek wie ilu dokładnie spośród trzech studentów zdało egzamin?</p>
4	<p><i>Fruits</i></p> <p>Mongines are small fruits that almost always have dried-out pith inside. Monica bought three mongine fruits yesterday.</p> <p>How many of the three fruits do you think have dried-out pith?</p> <p>Monica tells you on the phone: "I have looked at one/two/three of the three fruits. One/two/three fruits have dried-out pith."</p> <p>Now how many of the three fruits do you think have dried-out pith?</p> <p>Do you think Monica knows exactly how many of the three fruits have dried-out pith?</p>	<p><i>Owoce</i></p> <p>Jagody Goi to małe owoce, które prawie zawsze mają we wnętrzu wyschnięty miąższ. Monika zakupiła trzy jagody Goi wczoraj.</p> <p>Jak myślisz, ile spośród trzech jagód Goi ma wyschnięty miąższ?</p> <p>Monika informuje Cię przez telefon: "Sprawdziłam jedno/dwa/trzy spośród trzech owoców. Jeden/dwa/trzy owoce mają wyschnięty miąższ."</p> <p>Jak myślisz, ile spośród trzech owoców ma wyschnięty miąższ?</p> <p>Czy uważasz, że Monika wie ile dokładnie spośród trzech owoców ma wyschnięty miąższ?</p>
5	<p><i>Phones</i></p> <p>Broken Sigo phones almost always have burned-out transistors. Ben must repair three broken Sigo phones.</p> <p>How many of the three phones do you think have burned-out transistors?</p> <p>Ben tells you on the phone: "I have looked at two of the three phones. Some of the phones have burned-out transistors."</p> <p>Now how many of the three phones do you think have burned-out transistors?</p> <p>Do you think Ben knows exactly how many of the three phones have burned-out transistors?</p>	<p><i>Telefony</i></p> <p>Zepsute telefony marki Sigo prawie zawsze mają przepalone tranzystory.</p> <p>Bartosz musi naprawić trzy telefony marki Sigo.</p> <p>Jak myślisz, ile spośród trzech telefonów marki Sigo ma przepalone tranzystory?</p> <p>Bartosz informuje Cię przez telefon: "Sprawdziłem dwa spośród trzech telefonów. Niektóre telefony mają przepalone tranzystory."</p> <p>Jak myślisz, ile spośród trzech telefonów ma przepalone tranzystory?</p> <p>Czy uważasz, że Bartosz wie ile dokładnie spośród trzech telefonów ma przepalone tranzystory?</p>

Table 1: (continued)

Nr.	English version	Polish version
6	<i>Letters</i> Letters to Laura’s company almost always have cheques inside. Today Laura received three letters. How many of the three letters do you think have cheques inside? Laura tells you on the phone: “I have looked at two of the three letters. Some of the letters have cheques inside.” Now how many of the three letters do you think have cheques inside? Do you think Laura knows exactly how many of the three letters have cheques inside?	<i>Listy</i> Listy do Laury prawie zawsze mają czeki w środku. Dzisiaj Laura otrzymała trzy listy. Jak myślisz, ile spośród trzech listów ma czeki w środku? Laura informuje Cię przez telefon: “Sprawdziłam dwa spośród trzech listów. Niektóre listy mają czeki w środku.” Jak myślisz, ile spośród trzech listów ma czeki w środku? Czy uważasz, że Laura wie ile dokładnie spośród trzech listów ma czeki w środku?

Corrections for multiple comparisons

Table 2: Corrections for multiple comparisons, experiment 1 (Gaetano 2013).

α 0.05	p	Rank (j)	p'	Outcome
	<0.001	1	0.001	SIG
	<0.001	2	0.001	SIG
	0.007	3	0.035	SIG
	0.021	4	0.084	NON SIG
	0.038	5	0.114	NON SIG
	0.290	6	0.580	NON SIG
	0.382	7	0.580	NON SIG

Table 3: Corrections for multiple comparisons, experiment 2 (Gaetano 2013).

α 0.05	p	Rank (j)	p'	Outcome
	<0.001	1	0.001	SIG
	<0.001	2	0.001	SIG
	<0.001	3	0.001	SIG
	<0.001	4	0.001	SIG
	<0.001	5	0.001	SIG
	<0.001	6	0.001	SIG
	0.296	7	0.592	NON SIG
	0.472	8	0.592	NON SIG

Experiment 1 results without the filtering of data with the control question about the speaker's knowledge

A one way ANOVA, in which the dependent measure were bets on 3 and the independent measure were scenarios (6 possibilities), showed no significant effect of scenario ($F(5) = 1.16$, $p = 0.327$). Thus, different scenarios did not affect the answers, so we will not include scenario effects in the analyses.

To evaluate the overall effect of access, we performed an ANOVA with access and word as independent measures and bets on 3 as the dependent measure. We found a main effect of access ($F(2) = 5.13$, $p = 0.006$, $\eta^2 = 0.034$), no interaction between word and access ($F(1) = 0.23$, $p = 0.631$, $\eta^2 = 0.001$ (in the original study $p < 0.001$)), and a main effect of word ($F(2) = 84.86$, $p < 0.001$, $\eta^2 = 0.371$).

Just as in the original study, we next used comparisons to test whether knowledge inferences were drawn (only) when predicted. We found such inference in the complete-access conditions: when the speaker said “two,” bets on state 3 were smaller than on state 2 (paired directional t-test, $t(48) = 6.29$, $p < 0.001$). The effect size was $d = 0.90$. When the speaker said “one,” bets on state 1 were bigger than on state 3 (paired directional t-test, $t(48) = 8.74$, $p < 0.001$). The effect size was $d = 1.25$. Moreover, when the speaker said “one,” bets on state one were also bigger than on state 2 (paired directional t-test, $t(48) = 9.838$, $p < 0.001$). The effect size was $d = 1.41$.

In contrast, there was no upper-bound of the numeral term when access was 1 and the speaker said “one”; bets on 1 were not greater than on 2 (paired directional t-test, $t(48) = 1.25$, $p = 0.216$, $d = 0.18$).

When access was one and the speaker said “one,” bets on 3 were also not bigger than on 1 (paired directional t-test, $t(48) = 1.58$, $p = 0.121$, $d = 0.23$).

Contrary to predictions, there was an upper bound of the numeral term when access was 2 and the speaker said “two”; bets on 2 were bigger than on 3 (paired, directional t-test, $t(48) = 5.28$, $p < 0.001$, $d = 0.76$). Note that the mean bets on prior probability distribution were not as predicted: on zero objects having the property = 4.29, one = 17.73, two = 46.76, three = 31.22.

When access was 2 and the speaker said “one,” we found the predicted weaker upper bound of the numeral term: bets on state 1 were significantly greater than on state 3 (paired directional t-test, $t(48) = 3.89$, $p < 0.001$, $d = 0.56$) as well as, surprisingly, on state 2, namely bets on 2 were not greater than bets on one (paired directional t-test, $t(48) = -0.855$, $p = 0.397$, $d = 0.12$).

Experiment 2 results without the filtering of data with the control question about the speaker's knowledge

A one way ANOVA, in which the dependent measure were bets on 3 and the independent measure were scenarios (6 possibilities), showed no significant effect of scenario ($F(5) = 1.85$, $p = 0.104$). Thus, different scenarios did not affect the answers, so we will not include scenario effects in the analyses.

To evaluate the overall effect of access, we performed an ANOVA with access and word as independent measures and bets on 3 as the dependent measure. We found a main effect of access ($F(2) = 46.82$, $p < 0.001$, $\eta^2 = 0.245$), an interaction between word and access ($F(1) = 33.38$, $p < 0.001$, $\eta^2 = 0.104$ (in the original study $p < 0.001$)), and a main effect of word ($F(2) = 190.81$, $p < 0.001$, $\eta^2 = 0.570$).

Just as in the original study, we next used comparisons to test whether knowledge inferences were drawn (only) when predicted. We found such inference in the complete-access conditions: when the speaker said “two,” bets on state 3 were smaller than on state 2 (paired directional t-test, $t(48) = 19.53$, $p < 0.001$). The effect size was $d = 2.79$. When the speaker said “one,” bets on state 1 were bigger than on state 3 (paired directional t-test, $t(48) = 10.00$, $p < 0.001$). The effect size was $d = 1.43$. Moreover, when the speaker said “one,” bets on state one were also bigger than on state 2 (paired directional t-test, $t(48) = 15.77$, $p < 0.001$). The effect size was $d = 2.25$.

In contrast, there was no upper-bound of the numeral term when access was 1 and the speaker said “one”; bets on 1 were not greater than on 2 (paired directional t-test, $t(48) = -0.03$, $p = 0.976$, $d = 0.004$).

Moreover, when access was one and the speaker said “one,” bets on 3 were bigger than on 1 (paired directional t-test, $t(48) = -2.18$, $p = 0.034$, $d = 0.31$).

As predicted, there was no upper bound of the numeral term when access was 2 and the speaker said “two”; bets on 2 were not bigger than on 3 (paired, directional t-test, $t(48) = 0.59$, $p = 0.557$, $d = 0.09$). Note that the mean bets on prior probability distribution were as predicted: on zero objects having the property = 4.96, one = 11.31, two = 31.53, three = 52.20.

When access was 2 and the speaker said “one,” we found the predicted weaker upper bound of the numeral term: bets on state 1 were significantly greater than on state 3 (paired directional t-test, $t(48) = 4.59$, $p < 0.001$, $d = 0.66$) as well as, bets on 2 were greater than bets on one (paired directional t-test, $t(48) = -3.00$, $p = 0.004$, $d = 0.43$).

References

- Atlas, Jay David & Stephen C. Levinson. 1981. It-Clefts, informativeness and logical form: Radical pragmatics (Revised Standard Version). In Peter Cole (ed.), *Radical pragmatics*. New York: Academic Press.
- Baratgin, Jean. 2009. Updating our beliefs about inconsistency: The Monty-Hall case. *Mathematical Social Sciences* 57(1). 67–95.
- Baratgin, Jean & Guy Politzer. 2006. Is the mind Bayesian? The case for agnosticism. *Mind & Society* 5(1). 1–38.
- Barnett, Vic. 1994. Outliers in sample surveys. *Journal of Applied Statistics* 21(5). 373–381.
- Baron, Jonathan. 2006. *Thinking and deciding*, 4th edn. Cambridge: Cambridge University Press.
- Bergen, Leon, Roger Levy & Noah Goodman. 2016. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics* 9. <https://doi.org/10.3765/sp.9.20>; <http://semprag.org/article/view/sp.9.20> (accessed 29 January 2018).
- Breheny, Richard. 2008. A new look at the semantics and pragmatics of numerically quantified noun phrases. *Journal of Semantics* 25(2). 93–139.
- Carston, Robyn. 1998. Informativeness, relevance and scalar implicatures. In Robyn Carston & Seiji Uchida (eds.), *Relevance theory: Applications and implications*, 179–236. Amsterdam: John Benjamins.
- Chierchia, Gennaro. 2004. Scalar implicatures, polarity phenomena and the syntax/pragmatics interface. In Adriana Belletti (ed.), *Structures and beyond: The cartography of syntactic structures*, 39–103. Oxford: Oxford University Press.
- Chierchia, Gennaro. 2006. Broaden your views: Implicatures of domain widening and the “logicality” of language. *Linguistic Inquiry* 37(4). 535–590.
- Chierchia, Genaro, Benjamin Spector & Danny Fox. 2011. The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In Paul Portner, Claudia Maienborn & Klaus von Stechow (eds.), *Handbook on semantics*. Berlin: De Gruyter.
- Degen, Judith, Michael Henry Tessler & Noah D. Goodman. 2015. Wonky worlds: Listeners revise common ground when utterances are odd. In Rick Dale, Carolyn Jennings, Paul P. Maglio, Teenie Matlock, David C. Noelle, Anne Warlaumont & Jeff Yoshimi (eds.), *Proceedings of the 37th annual conference of the Cognitive Science Society*, vol. 2, 548–553.
- Dieuleveut, Anouk, Emmanuel Chemla & Benjamin Spector. 2019. Distinctions between primary and secondary scalar implicatures. *Journal of Memory and Language* 106. 150–171.
- Evans, Jonathan St. B. T., Simon J. Handley, David E. Over & Nicholas Perham. 2002. Background beliefs in Bayesian inference. *Memory & Cognition* 30(2). 179–190.
- Fox, Danny. 2007. Free choice and the theory of scalar implicatures. In Uli Sauerland & Penka Stateva (eds.), *Presupposition and implicature in compositional semantics*, 71–120. London: Palgrave Macmillan UK. https://doi.org/10.1057/9780230210752_4; http://link.springer.com/10.1057/9780230210752_4 (accessed 19 November 2018).
- Frank, Michael C. & Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science* 336(6084). 998.
- Gaetano, Justin. 2013. *Holm-Bonferroni sequential correction: An EXCEL calculator - Ver. 1.2*. <https://doi.org/10.13140/RG.2.1.3920.0481>.
- Geurts, Bart. 2010. *Quantity implicatures*. Cambridge; New York: Cambridge University Press.
- Goodman, Noah D. & Michael C. Frank. 2016. Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences* 20(11). 818–829.

- Goodman, Noah D. & Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science* 5(1). 173–184.
- Grice, Paul. 1975. Logic and conversation. In Peter Cole & Jerry L Morgan (eds.), *Syntax and semantics*. New York: Academic Press.
- Grice, Paul. 1989. *Studies in the way of words*. Cambridge, Mass.: Harvard University Press.
- Holm, Sture. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6(2). 65–70.
- Horn, Laurence R. 1972. *On the semantic properties of logical operators in English*. Los Angeles, CA: University of California Doctoral Dissertation.
- Horn, Laurence R. 2006. Implicature. In *The handbook of pragmatics*. Malden, MA; Oxford: Blackwell. <https://doi.org/10.1002/9780470756959> (accessed 2 May 2015).
- Koch, Steffen. 2020. Engineering what? On concepts in conceptual engineering. *Synthese*. <https://doi.org/10.1007/s11229-020-02868-w>; <http://link.springer.com/10.1007/s11229-020-02868-w> (accessed 25 November 2020).
- Krynski, Tevya R. & Joshua Tenenbaum. 2003. The role of causal models in statistical reasoning. Richard Alterman, David Kirsch (eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society, Erlbaum*, 693–698.
- Levinson, Stephen C. 2000. *Presumptive meanings: The theory of generalized conversational implicature (Language, Speech, and Communication)*. Cambridge, Mass: MIT Press.
- Oaksford, Mike & Nick Chater. 2009. Précis of Bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences* 32(1). 69–84.
- Russell, Benjamin. 2006. Against grammatical computation of scalar implicatures. *Journal of Semantics* 23(4). 361–382.
- Sauerland, Uli. 2004. Scalar implicatures in complex sentences. *Linguistics and Philosophy* 27(3). 367–391.
- Sauerland, Uli. 2012. The computation of scalar implicatures: Pragmatic, lexical or grammatical?: Computation of scalar implicatures. *Language and Linguistics Compass* 6(1). 36–49.
- Scontras, Gregory, Michael Henry Tessler & Michael Franke. 2018. Probabilistic language understanding: An introduction to the Rational Speech Act framework. Available at: <https://www.problang.org>.
- Skoczeń, Izabela & Aleksander, Smywiński-Pohl. *Quantifier term scope ambiguities in uncertainty contexts strengthen scalar implicature*. Mouton de Gruyter, manuscript.
- Skoczeń, Izabela & Aleksander, Smywiński-Pohl. The context of mistrust: Perjury ascriptions in the courtroom. In Laurence Horn, Stein Dieter, Janet Giltrow (eds.), *From lying to perjury: Linguistic and legal perspectives on lies and other falsehoods, foundations of language and law*. Mouton de Gruyter. Mouton de Gruyter, forthcoming 2022.
- Spector, Benjamin. 2003. Scalar implicatures: Exhaustivity and Gricean reasoning. In Balder Ten Cate (ed.), *Proceedings of the eighth ESSLLI student session*, 277–88. Vienna: FOLLI.
- Spector, Benjamin. 2013. Bare numerals and scalar implicatures. *Language and Linguistics Compass* 7(5). 273–294.
- Sperber, Dan & Deirdre Wilson. 2006. Relevance theory. In Laurence R. Horn & Gregory L. Ward (eds.), *The handbook of pragmatics*. Malden, MA; Oxford: Blackwell. <https://doi.org/10.1002/9780470756959> (accessed 2 May 2015).
- Stanovich, Keith E. & Richard F. West. 1998. Individual differences in rational thought. *Journal of Experimental Psychology: General* 127(2). 161–188.

Yoon, Erica, Michael Henry Tessler, Noah D. Goodman & Michael C. Frank. 2020. Polite speech emerges from competing social goals. *Open Mind: Discoveries in Cognitive Science* 4. 71–87.

Bionotes

Izabela Skoczeń

Jagiellonian University, Krakow, Poland
izabela.skoczen@uj.edu.pl

Izabela Skoczeń specializes in experimental studies of language, morality and law. She applies the methodology of cognitive science to shed new light in philosophical debates. She published a number of articles in international journals as well as a book “Implicatures within Legal Language” (2019). Her research focuses on inferences in uncertainty contexts. She is a member of the Guilty Minds Lab at the University of Zurich and the Jagiellonian Centre for Law, Language and Philosophy in Krakow.

Aleksander Smywiński-Pohl

AGH University of Science and Technology, Krakow, Poland

Aleksander Smywiński-Pohl is a researcher in natural language processing. He received his Ph.D. in 2015 from AGH University of Science and Technology in Krakow for the work: Automatic extraction of semantic relations from Polish texts. His research concentrates on the application of modern NLP techniques in a broad range of practical problems. In 2017, he started a project funded by the Polish National Center for Research and Development for the construction of an intelligent legal information system called “Lemkin”.