Proceedings of SALT 33: 439–459, 2023

# On the meaning of intonational contours: a view from scalar inference[*]

Alexander Göbel
*Princeton University*

Eszter Ronai
*Northwestern University*

**Abstract** This paper investigates the meaning of intonational contours by experimentally testing how they affect the likelihood of scalar inference (SI) calculation. Our main test case is the rise-fall-rise contour (RFR) which, based on prior theoretical work, is predicted to either increase or decrease the likelihood of SI. We conducted two experiments using an inference task: one where participants first produce a target sentence with their choice of contour and one where participants listen to a pre-recorded target sentence with a particular contour. The experiments converged in showing that the RFR increases SI rate relative to a neutral fall. Additionally, production data revealed the frequent use of another contour that resembles the Contradiction Contour, which we label Concession Contour. This contour also led to an increase in SI rate, although to a lesser extent than the RFR. In addition to informing the theoretical literature on RFR, our results also highlight the methodological importance of controlling for intonation in the study of SI.

**Keywords:** contradiction contour, intonation, production, rise-fall-rise, scalar diversity

## 1 Introduction

Intonation constitutes an essential component of the meaning of an utterance, for example in the form of stress placement for question-answer congruence (Jackendoff 1972) or in the way prosodic phrasing affects scope (Hirotani 2005). Here we focus on intonational contours and the effect they have on the interpretation of an utterance. The main focus of this paper is the so-called rise-fall-rise contour (RFR), exemplified by the underlined parts of (1), which has been analyzed in terms of uncertainty (Ward & Hirschberg 1985) and related notions.

(1) *CK*: If everybody knew everybody, we wouldn't have the problems we have in the world today. You don't rob somebody if you know their name.

[*] We are indebted to Emma Nguyen and Luke Adamson for providing audio stimuli, as well as Dan Goodhue, Sunwoo Jeong, Deniz Rudin, Michael Wagner, the UPenn Experimental Semantics Lab, and the SALT audience for feedback. This material is partially based upon work supported by the National Science Foundation under Grant No. #BCS-2041312. Authors contributed equally to this work and are listed in alphabetical order.

*JS*: You're robbin' me... (AUDIO)

One empirical domain where different theoretical accounts of the RFR make different predictions concerns its effect on scalar inferences (SIs), which this paper tests in a production and a perception experiment. Specifically, we will assess the RFR - and intonational contours more generally - in the context of scalar diversity, which is the observation that the likelihood of SI calculation varies substantially depending on the lexical scale (van Tiel, Van Miltenburg, Zevakhina & Geurts 2016). This approach thus makes it possible to relate intonation to a broader range of lexical meanings and to assess the role intonation in scalar diversity, a domain where prior studies have only used written materials.

In addition to the RFR, another contour that will be discussed is what we we term the Concession Contour, illustrated in (2). This contour was frequently used in the production experiment. It prosodically resembles the so-called Contradiction Contour (Liberman & Sag 1974), illustrated in (3), but intuitively makes a different contribution. We will provide further discussion in Section 4.

(2) *JS*: Well, it's only a year. That's not so bad. (AUDIO)

(3) *JS*: These balloons aren't gonna stay filled 'til New Year's!
*CK*: Those aren't for New Year's! Those are my everyday balloons. (AUDIO)

The structure of this paper is as follows. Section 2 provides background on previous research on the RFR and scalar diversity, turning finally to what predictions we can take different accounts of the RFR to make with respect to its effect on SI calculation. Section 3 presents the two experiments. Section 4 discusses the implications of the results and concludes the paper.

## 2 Background

### 2.1 RFR

An early influential account of the RFR comes from Ward & Hirschberg (1985) (henceforth W&H), who propose that the RFR conveys speaker uncertainty with respect to a scale. The authors primarily focus on the RFR in replies to questions as in (4), where its contribution can be intuitively described as a polite hedge. W&H capture data like this by proposing that the RFR conveys uncertainty either about whether it is appropriate to evoke a scale (4a), what scale is being evoked (4b), or where a particular value falls on a given scale (4c).

(4) a. A: Are you leaving today?
B: I'm not leaving TODAY... Ward & Hirschberg: (54)

b. A: Are you a doctor?
B: I have a PHD... Ward & Hirschberg: (58)

c. A: Have you ever been West of the Mississippi?
B: I've been to MISSOURI... Ward & Hirschberg: (62)

As an alternative but related proposal, Constant (2012) argues that the RFR quantifies over assertable alternatives and indicates that they cannot be claimed by the speaker, formalized in (5).

(5) $[\![\text{RFR } \phi]\!]^{ci}$ =
$\forall p \in [\![\phi]\!]^f$ s.t. $p$ is assertable in $C$: the speaker cannot safely claim $p$.

One pattern this account is designed to capture is that the RFR can only occur when the alternatives to the stressed element do not resolve all other alternatives, illustrated in (6). Both maximal scale elements, which either entail the falsity of all stronger alternatives (*no one*) or entail the truth of all weaker alternatives (*all*), are infelicitous, while the element that leaves alternatives open (*most*) is not.

(6) A: Did your friends like the movie?

a. B: Most of my friends liked it...

b. B: #No one liked it...

c. B: #All of my friends liked it... Constant: (33)-(34)

Further related accounts come from Wagner (2012) and Wagner, McClay & Mak (2013). Wagner differs from Constant in assuming that the RFR operates over speech acts, formalized in (7), presupposing that a salient alternative is possibly true. This adjustment is meant to capture the RFR's ability to be embedded, as with the appositive relative clause in (8).

(7) $[\![\text{RFR}]\!] = \lambda S.\ \exists S' \textit{ in } [\![S]\!]^{g}{}_{a}, S \not\twoheadrightarrow S'$ and performing $S'$ might be justified: $S$

(8) John - <u>who likes sweets</u> - was an obvious suspect.

Wagner et al. focus on the incompleteness component of the RFR, stated in (9), and present experimental evidence that the RFR is produced more frequently and perceived as more acceptable in partial answers, compared to complete answers.

(9) **RFR** ($p$): The speaker asserts $p$ but considers it to be only an incomplete answer to the question under discussion.

Although the previous three accounts seem closely related, they differ in a small but important detail. While all three accounts are compatible with the RFR providing an incomplete answer when the truth of other alternatives is unknown, Constant additionally allows alternatives to be unclaimable because they are known to be false. This feature captures the fact that the RFR can be followed up with an answer that fully resolves the relevant question, as in (10), which is incompatible with Wagner and Wagner et al.'s accounts.

(10) A: Did your friends like the movie?
B: JOHN liked it... the rest of them hated it. Constant: (16)

A different account comes from Westera (2019), which can be viewed as elaborating on the relevance of the question under discussion (QUD, Roberts 2012) that Wagner et al. highlighted. Westera proposes that the RFR - assumed to also cover cases of Contrastive Topic (Büring 2003) - indicates that a maxim is suspended relative to the main QUD while a secondary QUD is being compliantly addressed. The infelicity of exhaustive answers with the RFR as in (6b) and (6c) is then captured because the maxims regarding the main QUD are being adhered to.

Lastly, Göbel (2019) and Göbel & Wagner (2023) shift their attention to the function of the RFR in argumentative dialogues. The observation they make is that the RFR exhibits an asymmetry in replies to statements depending on the "polarity" of the initial statement, which they dub valence asymmetry. While the RFR is felicitous when providing a positive counterpoint to a negative statement (11a), it is degraded when the order is reversed and its carrier utterance provides a negative counterpoint to a positive statement (11b).

(11) a. A: The bike ride yesterday was really terrible, the weather was horrific.
B: We had a cocktail... (AUDIO)
b. A: The bike ride yesterday was really great, the weather was perfect.
B: #We had an accident... (AUDIO)

Crucially, this pattern is unexplained by previous accounts insofar as B's replies in both (11a) and (11b) do not differ in whether alternatives are left open or not. The authors hence propose that the RFR conveys the presence of a non-entailed stronger alternative on a pragmatically inferred scale, formalized as in (12).

(12) $[\![\text{RFR}]\!](Q_{<s,<<s,t>,t>})(p_{<s,t>})(w)$:
$\exists q[q \in Q(w)\ \&\ p \nRightarrow q \wedge q < p \wedge q(w)].\ p(w)$

For cases like (11), this scale concerns an evaluation, here of the quality of the bike ride, where the positive reply implies a stronger, or better, alternative to A's statement, whereas the negative reply implies a weaker, or worse, one. For cases like (6), on the other hand, the scale is one of logical entailment such that a stronger alternative to *most* would be *all*, capturing the pattern in a similar way as previous accounts.

The next subsection provides background on studies of SI, scalar diversity, and existing work testing the role of intonation in modulating SI calculation.

## 2.2 SI and scalar diversity

SI represents one of the classic examples of pragmatic enrichment. An utterance containing the quantifier *some*, for example, is often enriched to mean *some but not all* —see (13).

(13) Fatima caught some of the mice.

a. Fatima caught at least some of the mice. literal

b. Fatima caught some, but not all, of the mice. SI-enriched

While there are many different theoretical proposals as to how SIs arise, a standard (neo-)Gricean account posits the following. Hearers assume that speakers are following the Maxim of Quantity (Grice 1967), and are therefore trying to be as informative as is required in the context. A more informative alternative utterance to (13) would have been *Fatima caught all of the mice*, where informativity can be defined as asymmetric entailment (Horn 1972). Observing that this more informative, stronger alternative was not uttered, hearers can then reason that it must be false, and the speaker chose not to utter it in order to avoid violating the Maxim of Quality. This leads them to derive the negation of the unsaid alternative which, combined with the original utterance's literal meaning (13a), results in the SI-enriched meaning (13b).

While the *some but not all* SI, based on the <*some*, *all*> lexical scale, is the most widely discussed example, SI can also arise from other pairs of lexical items that form a scale. The example in (14), for instance, is based on the <*happy*, *ecstatic*> scale.

(14) The winner is happy.

a. The winner is at least happy. literal

b. The winner is happy, but not ecstatic. SI-enriched

Hearers of the weaker utterance in (14) reason that the speaker did not utter the more informative alternative *The winner is ecstatic* because it would not have been true. The weaker utterance's literal meaning (14a) and the negation of the stronger alternative together give rise to the SI-enriched meaning (14b). But while the mechanism underlying these two different SIs is posited to be the same, the likelihood of different lexical scales leading hearers to derive SI in fact varies substantially. In van Tiel et al.'s (2016) highly influential study, the rate at which participants calculated SIs ranged from 4% to 100% depending on the scale (see also earlier work by Beltrama & Xiang 2013; Baker, Doran, McNabb, Larson & Ward 2009; Doran, Ward, Larson, McNabb & Baker 2012). This variation has been termed *scalar diversity*.

Studies following van Tiel et al. (2016) have concentrated on answering the question of what can explain the observed inter-scale variation in SI calculation. How likely a scale is to lead to SI has been related to various properties of the stronger alternative, or of the relationship between the weaker scalar term and that alternative (van Tiel et al. 2016; Gotzner, Solt & Benz 2018; Ronai & Xiang 2022b; Hu, Levy & Schuster 2022; Hu, Levy, Degen & Schuster 2023; Westera & Boleda 2020). Studies have also suggested that propensity for SI is linked to another type of semantic process or pragmatic inference that is variable across scales (Gotzner et al. 2018; Sun, Tian & Breheny 2018). Yet other work has investigated the role of context, or contextual relevance, in explaining scalar diversity (Simons & Warren 2018; Pankratz & van Tiel 2021; Ronai & Xiang 2021). One shortcoming of this existing body of work that we would like to highlight, however, is that all prior studies on scalar diversity have used exclusively written experimental stimuli, or modeled data from other studies that had done so. This is despite the fact that - as we will review below - intonation is known to affect SI calculation more generally. Thus there is reason to believe that using auditory stimuli, carefully controlling and manipulating the intonation with which SI-triggering utterances are produced, could uncover interesting patterns that written studies on scalar diversity have obscured.

As mentioned, there are robust findings in the literature showing that intonation affects how likely SI is to arise. Several studies in this domain have focused on ad hoc scales (Hirschberg 1985) giving rise to exhaustive inferences. For example, in a mouse-tracking experiment, Tomlinson & Ronderos (2021) investigated the exhaustive interpretation arising from dialogues such as (15).

(15) A: Were Manu and Moni at the party?
B: Manu was there.

The authors were interested in the derivation of the inference that Speaker B believes that Manu was there at the party, but Moni was not (= Speaker B believes that (¬Moni, Manu)). They compared B's utterance when pronounced with the L+H* vs. L*+H contour in German and found that SI derivation rates were both lower and more delayed with the L+H* contour. While this finding is surprising in the sense that the L*+H contour is the one taken to index uncertainty, it nevertheless constitutes evidence that intonation significantly affects likelihood of SI calculation. A similar conclusion can be drawn from Gotzner (2019), who found that participants computed more exhaustive inferences in German with an L+H*, as compared to an H* accent in a truth value judgment task. For another comparison of the effect of L+H* vs. H* on ad hoc SIs, see John M. Tomlinson, Gotzner & Bott (2017), who showed that the inference is processed earlier under the former contour.

There exists also some work testing not ad hoc, but lexical scales (Horn 1972; Levinson 2000). Using truth value judgements, Chevallier, Noveck, Nazir, Bott, Lanzetti & Sperber (2008) showed that prosodic stress on *or* resulted in an increase in exclusive *not both* interpretations: an SI based on the <*or*, *and*> scale. Most recently, Buccola & Goodhue (to appear) have looked at the role of intonation for SIs arising from the <*some*, *all*> scale, testing dialogues such as (16).

(16) A: Did Bonnie eat all of the pears?
B: Bonnie ate some of the pears.

B's answer was manipulated such that it was either pronounced with a falling contour (H* L-L%) or the RFR (L*+H L-H%). In Experiment 2, participants were instructed to pair each of the contours with either an SI (=B thinks that Bonnie didn't eat all of the pears) or ignorance inference (=B isn't sure whether or not Bonnie ate all of the pears) interpretation. Results revealed that participants were significantly more likely to associate the fall with an SI and the RFR with an ignorance inference than the other way around. While this study did not directly test whether SI rates increase or decrease with particular contours, it does add to the body of evidence demonstrating that SI-related interpretations are sensitive to intonational cues.

As mentioned above, despite well-documented effects of intonation on SI calculation, work on scalar diversity has tended to use written stimuli. Nonetheless, there are two notable exceptions, that is, two studies that manipulated intonation while testing multiple different lexical scales, which we now turn to. Most relevantly, de Marneffe & Tonhauser (2019) tested 16 different adjectival scales in an experiment where participants were presented with dialogues like (17).

(17) Mike: Was your hike exhausting?
Julie: It was strenuous.

The authors manipulated whether Julie's answer was pronounced with a neutral (H* L-L%) or RFR (L*+H L-H%) contour, and found that the RFR made SI interpretations (e.g., *strenuous but not exhausting*) more likely. However, this study does not report by-item results, leaving open the question of how (or whether) the intonation manipulation interacted with scalar diversity.

Cummins & Rohde (2015) tested 20 different adjectival scales, and presented participants with sentences such as *The view from the hotel room is pretty* in two intonation conditions: neutral vs. with focus placement on the scalar adjective (here, *pretty*). The authors take the focus manipulation to be a manipulation of the QUD, which they predict would influence SI rates. Indeed, they found that participants were more likely to calculate the SI (e.g., *not gorgeous*) in the focus condition. However, as their by-item results (p. 7, Figure 1) show, scalar terms differ in how susceptible they are to the intonation manipulation. There is substantial variation in effect size - i.e., in how much more likely the SI was to be calculated in the focus condition than in the neutral condition - and 6 scales in fact show the opposite pattern to the overall effect. This suggests that it is indeed important to study the effects of intonation on SI calculation across many scales, and to study scalar diversity with auditory stimuli. Crucially, one way in which our study differs from Cummins & Rohde (2015) is that we are interested in more complex intonational contours over the whole SI-triggering utterance (e.g., the RFR), rather than just manipulating whether the weaker scalar term is focused.

Next, we discuss what predictions might be derived from different theoretical accounts of the RFR for its potential effect on the likelihood of SI calculation.

### 2.3 The effect of the RFR on SI calculation

What predictions do these accounts make with respect to the RFR's effect on SI calculation, relative to neutral intonation with a declarative fall? Starting with Ward & Hirschberg (1985), there is an open question about what level the uncertainty could be conveyed at, given the different options provided. Following de Marneffe & Tonhauser (2019), we will assume that the most sensible option is one where uncertainty relates to the choice of scalar value rather than the existence or type of scale, given that the target items in studies of SI are inherently scalar. The examples of this type of uncertainty that Ward & Hirschberg discuss (their Type III) all involve a sense of the speaker proffering an answer as being potentially insufficient, as is the case in (4c) above. This usage may make it seem as if the speaker of the RFR is themselves not committed to their reply being a yes or no answer but is leaving it to the hearer to decide. Given that an SI requires the negation of stronger alternatives, we thus take this account to predict the RFR to decrease the rate of SIs drawn.

For Constant (2012), the situation is more complex given that alternatives can be unclaimable either because they are considered false or because they are not known. If the RFR is taken to indicate that stronger alternatives are false, we would expect an increase in SI rate. On the other hand, if the RFR conveys uncertainty regarding stronger alternatives - like Ward & Hirschberg (1985) - then we expect the opposite: that the RFR would decrease SI rate. Constant's account would thus be compatible with either outcome.

For Wagner (2012), Wagner et al. (2013) and Westera (2019), the predictions are more clear cut. In the case of Wagner, the relevant alternative has to be possibly true, which would not be the case if it is negated to draw an SI; Wagner et al. and Westera even mention SIs explicitly as something the RFR cancels. Westera also discusses the possibility, however, that the question the RFR-utterance replies to may not be taken as the main QUD, in which case it would be compatible with exhaustivity. However, given that it is not straightforward how to flesh out this possibility, we will take Westera's account to predict a decrease in SI rate as well.

Finally, Göbel (2019) and Göbel & Wagner (2023) simply treat the RFR as

| Account | Effect on SI Rate |
|---|---|
| Ward & Hirschberg (1985) | ⇓ |
| Constant (2012) | ? |
| Wagner (2012) | ⇓ |
| Wagner et al. (2013) | ⇓ |
| Westera (2019) | ⇓ |
| Göbel (2019), Göbel & Wagner (2023) | ⇑ |

**Table 1** Prediction for SI rate by accounts of RFR.

implying the existence of a stronger alternative while remaining agnostic regarding its truth value. On this account, a possible effect of the RFR could then be that highlighting the salience of the relevant alternative leads to an increase in SI rate. This view is supported by Ronai & Xiang (2022a), who found that a prior question that mentions the stronger alternative leads to an increase in SI rate, relative to when the SI-triggering sentence occurs without a question context, or following a question that mentions the weaker scalar term itself.

A summary of the discussed predictions is shown in Table 1. With this background in mind, we now turn to the presentation of the two experiments.

## 3 Experiments

The two experiments presented in this section aim to test the predictions the different accounts of the RFR make with respect to its effect on the rate of SI calculation. We conducted both a production- and a perception-oriented experiment in order to provide converging evidence across different methodologies.

### 3.1 Experiment 1: Production + Inference Task

#### 3.1.1 Method, Materials & Design

Stimuli consisted of question-answer dialogues containing scalar terms as in (18). These varied in whether the question prompt (Emma's question) and the target sentence (the participant's reply) contained the same weaker scalar term (18a) or the question contained the chosen stronger alternative (18b). There were 60 lexical scales taken from Ronai & Xiang (2022a) in addition to 20 fillers. Participants saw each item only in one condition (SAME vs. STRONG) in a Latin-square design.

(18) Sample Item, Experiment 1

a. *Emma*: Was the winner happy? SAME

b. *Emma*: Was the winner ecstatic? STRONG

*You*: She was happy.

*Given your response, do you think Emma would conclude that the winner is not ecstatic?*

Participants first saw the full dialogue on the screen. After pressing a button, they heard an audio recording of the question - which was included to make the task more natural - and had to record themselves saying the reply. Afterwards, they were given the task question *Given your response, do you think...?* (italicized in (18)) and chose between "Yes" and "No" as their answer. In this adapted version of the inference task from van Tiel et al. (2016) (see also i.a., Pankratz & van Tiel 2021), if a participant responds with "Yes", that can be taken to index SI calculation: that the participant has enriched *happy* to *not ecstatic*. Responding with "No", on the other hand, suggests that the participant has not calculated the SI and takes *happy* to be compatible with *ecstatic*. Altogether, this method allowed us to gather data on the production rates of relevant contours across conditions and items, as well as examine SI rates given that a certain contour was produced.

### 3.1.2 Procedure

The experiment was implemented through prosodyExperimenter (https://github.com/prosodylab/prosodylabExperimenter). Participants first saw a welcome screen, followed by a chance to adjust their volume and test their microphone, an online consent form, and a language background questionnaire. Afterwards, there was a test where participants were played three sounds and had to choose which one was the quietest, which required the use of headphones. For the main part of the experiment, participants provided their production of the target sentence and then answered the question for the inference task, as described above. There were three practice trials after receiving instructions, followed by a total of 80 stimuli. The experiment concluded with a chance to provide feedback. A test version of the experiment can be accessed at https://prosodylab.org/~agobel/conepi/30-scaRFR_Pro2AFC/?SESSION_ID=SALT&mode=experiment.

### 3.1.3 Participants

64 monolingual native speakers of American English were recruited on Prolific and compensated $4 or $5 (depending on time). One participant's response file was not properly saved, and 26 participants were excluded for providing unnaturally monotone or otherwise unusable recordings, leaving 37 participants for data analysis reported below.

### 3.1.4 Predictions

First, different theoretical accounts make different predictions for SI rates with the RFR relative to a neutral fall - these were shown in Table 1. Second, the following predictions can be made regarding production rates. One expectation that serves as a sanity check is that participants should produce Verum Focus (i.e., shift prominence to the auxiliary, Höhle 1992) in the SAME condition, since everything in the sentence is given. However, this is only true for items containing an auxiliary, which was not the case for all stimuli. Additionally, Göbel (2019) and Göbel & Wagner (2023) would predict the RFR to occur more frequently in the STRONG condition, given that the requirement for a stronger alternative to be present is explicitly satisfied.

### 3.1.5 Results

**Production rates** Recordings were manually annotated by the first author in terms of the overall contour used by the participant on a given item. The "a priori" categories originally included five contours: Neutral Fall, RFR, Verum Focus, Rising Declaratives, and Other/Unclear. However, after initial inspection, two changes were made. First, Rising Declaratives was taken out due to the contour not occurring sufficiently frequently. Second, as mentioned in Section 1, there was a notably

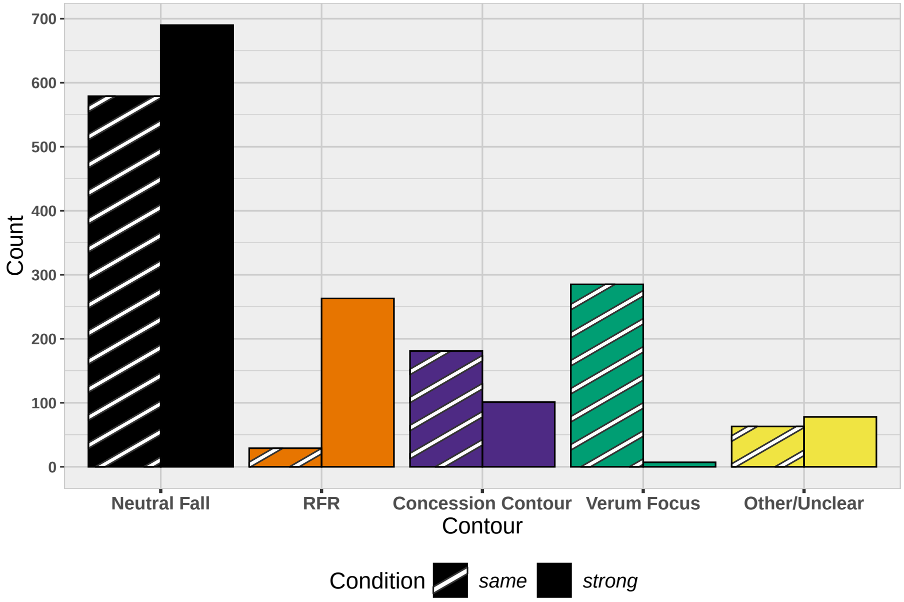

**Figure 1** Production rates by contour and condition.

frequent use of another contour that we labeled “Concession Contour”, so it was added as one of the categories. The counts by condition for each category are shown in Figure 1.

The first thing to note is that Neutral Fall is by far most frequent contour to be used, comprising about 56% of the total recordings even after excluding monotonous participants, indicating the difficulty of motivating participants to be more creative with their intonation choices in an online setting.[1] Next we can see that the RFR was almost exclusively used in the STRONG condition, in line with the prediction by Göbel & Wagner (2023) and Göbel & Wagner (2023). The Concession Contour, on the other hand, had a trend toward occurring more frequently in the SAME condition, but was more evenly distributed. Another prediction was borne out by the fact that Verum Focus virtually exclusively occurred in the SAME condition.

**SI rates by contour** We next looked at the rate of SI calculation from the inference task depending on the contour produced by the participant. We restricted this analysis to Neutral Fall as a baseline, RFR as the intended contour of interest, and the Concession Contour as the third most frequent contour for exploratory purposes.

[1] Note, however, that the Neutral Fall category included a lot of internal variation that goes beyond the shared overall pitch contour.

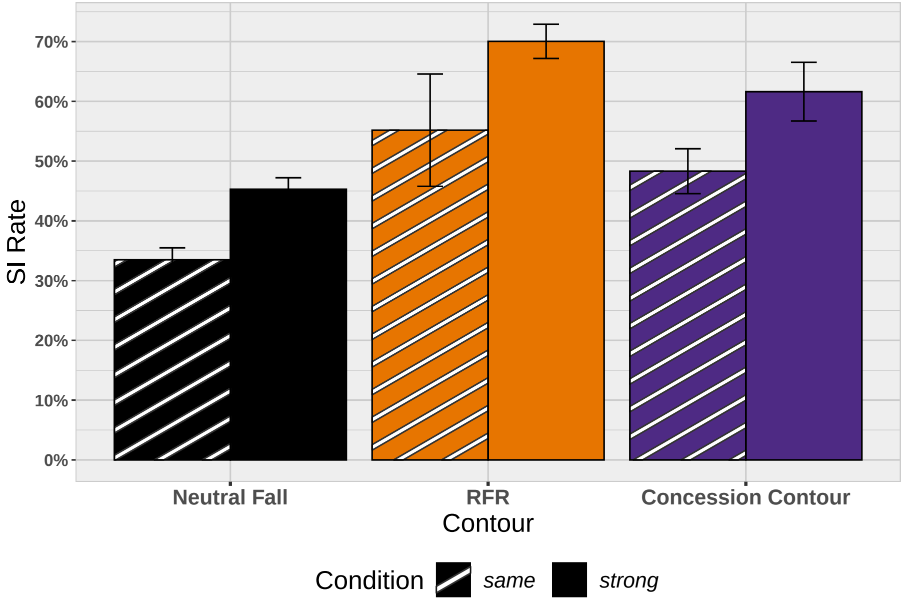

**Figure 2** SI rates by contour and condition.

SI rates for those three contours by condition are shown in Figure 2. For the statistical analysis, we fit a logistic mixed effects regression model using the lme4 package in R (Bates, Mächler, Bolker & Walker 2015). The model predicted Response ("Yes" vs. "No") as a function of Contour (RFR vs. Neutral Fall vs. Concession Contour), Strength (SAME vs. STRONG) and their interaction. It included the maximal random effects structure supported by the data (Barr, Levy, Scheepers & Tily 2013): random by-participant and by-item intercepts and slopes for the Strength predictor. Both fixed effects predictors were treatment-coded: in Contour, the Neutral Fall level served as baseline, while in Strength, the STRONG level served as baseline.

The statistical analysis revealed the following results. First, the SAME condition produced lower SI rates than the STRONG condition (Estimate=-1.13, SE=0.27, $z$=-4.14, $p$ <0.001), replicating Ronai & Xiang (2022a). There was no evidence that this effect differed across contours, i.e., there were no significant interactions (Estimate=-0.58, SE=0.59, $z$=-0.98, $p$ =0.33; Estimate=0.04, SE=0.4, $z$=0.11, $p$ =0.92). Second, Neutral Fall showed the lowest SI rate (33.5% in the SAME and 45.3% in the STRONG condition), followed by the Concession Contour (48.3% in the SAME and 61.6% in the STRONG condition), which produced a significantly higher rate (Estimate=0.7, SE=0.31, $z$=2.25, $p$ <0.05). Lastly, the RFR produced the highest SI rate (55.2% in the SAME and 70% in the STRONG condition), also significantly higher than the baseline Neutral Fall (Estimate=0.89, SE=0.23, $z$=3.81, $p$ <0.001).

### 3.1.6 Discussion

The experiment provided data from two sources: production rates of contours and inference rates given the production of a certain contour (Göbel 2019; Göbel & Wagner 2023). Production rates showed that the RFR was almost exclusively used when the question prompt mentioned a stronger alternative, in line with the prediction of the salience account. Additionally and more crucially, using the RFR led to an increase in SI rate relative to using a Neutral Fall. This finding goes against the predictions of Ward & Hirschberg (1985), Wagner (2012), Wagner et al. (2013), and Westera (2019), is compatible with Constant (2012), and supports Göbel (2019) and Göbel & Wagner (2023).

However, a possible objection to this conclusion is that the SI rates could be driven by lexical properties of the items. As discussed in Section 2.2, research on scalar diversity has revealed many factors that contribute to differences in SI rate, independent of intonation. Given the nature of the present task, it may simply be the case that the RFR was used more frequently with items - that is, lexical scales - that would show higher SI rates irrespective of intonation. While such a pattern would require its own explanation, the consequence for the present study would be that the observed differences in SI rate between intonational contours are merely an epiphenomenon. Indeed, looking at production rates across items, shown in Figure 3, there is clear variation regarding when the RFR was more likely to be used (which we will come back to later). To address these concerns, the next experiment used a perception task that allows assessing the contribution of intonation independently of lexical factors.

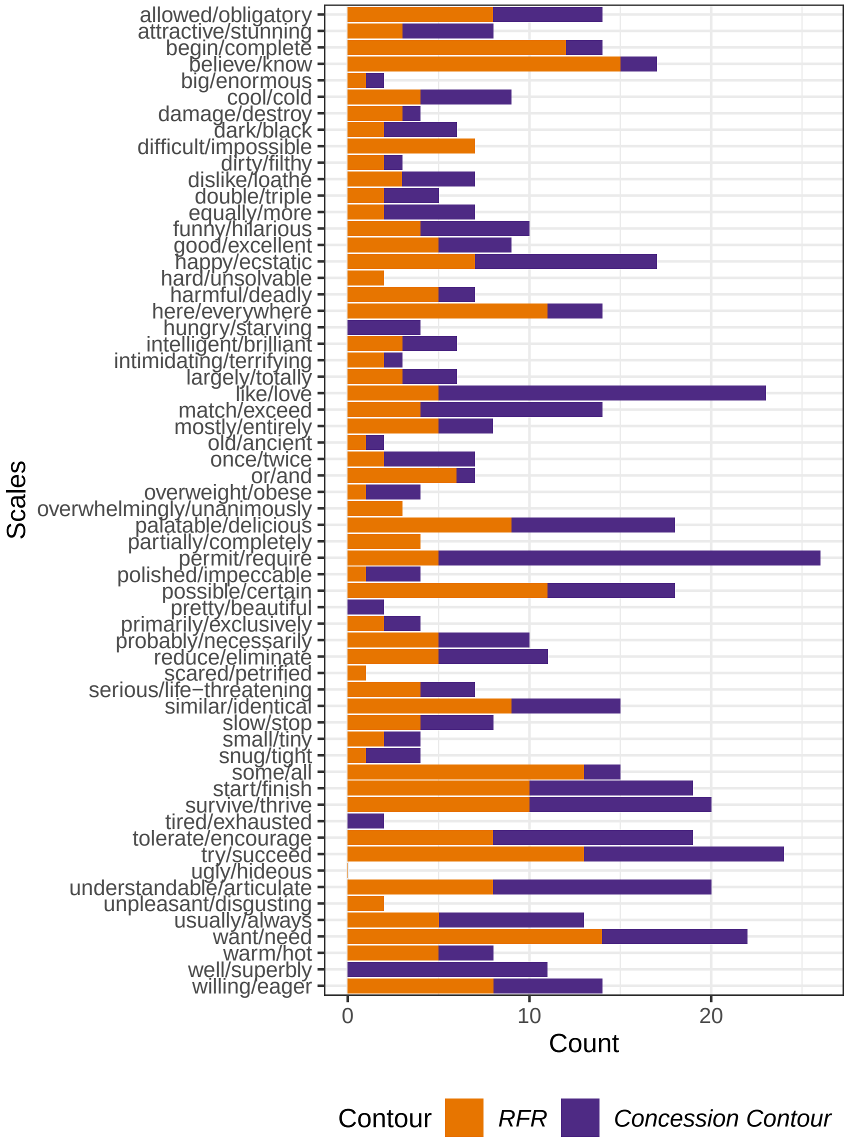

**Figure 3** Production rates for RFR and Concession Contour by item.

## 3.2 Experiment 2: Perception + Inference Task

### 3.2.1 Method, Materials & Design

We used the same materials as Experiment 1 (60 experimental stimuli + 20 fillers), but restricted to the STRONG condition, since the RFR was rarely produced in the SAME condition. Additionally, both the question prompt and the target sentence were presented auditorily without the text being visible on the screen, with the target sentence occurring with one of three contours: a NEUTRAL FALL, the RFR, or the CONCESSION CONTOUR, again in a Latin-square design. After listening to one version of the dialogue, participants were asked the same task question - with the only modification being that the target speaker was no longer referred to as *you* but as *Luke*. As before, we take a "Yes" response to index SI calculation, and a "No" response to index that the participant has not calculated the SI. A sample item with recordings is shown in (19).

(19) Sample Item, Experiment 2

*Emma*: Was the winner ecstatic?

*Luke*: She was happy. {[NEUTRAL], [RFR], [CONCESSION]}

*Given Luke's response, do you think Emma would conclude that the winner is not ecstatic?*

### 3.2.2 Procedure

The general procedure was largely the same as for Experiment 1, except there was no mic check. A test version can be accessed at https://prosodylab.org/~agobel/conepi/31-scaRFR_Aud2AFC/?SESSION_ID=SALT&mode=experiment.

### 3.2.3 Participants

90 monolingual native speakers of American English were recruited on Prolific and compensated $2.50. 17 participants were excluded for failing the headphone test. Data from the remaining 73 participants is reported below.

### 3.2.4 Results

SI rates by contour are shown in Figure 4. To analyze the results, we fit a logistic mixed effects regression model predicting Response ("Yes" vs. "No") as a function of Contour (Neutral Fall vs. RFR vs. Concession Contour). The fixed effects predictor was treatment coded, with Neutral Fall as the reference level. The maximal converging random effects structure included by-participant intercepts and by-item

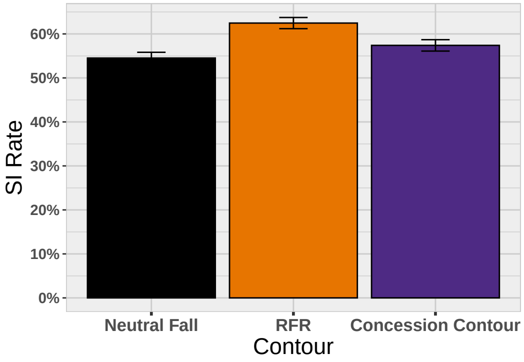

**Figure 4** SI rates by contour.

intercepts and slopes. We found significantly higher rates of SI calculation with the RFR than with the Neutral Fall (Estimate=0.4, SE=0.12, $z$ =3.25, $p$ <0.01). The difference between Neutral Fall and Concession Contour, on the other hand, was not significant (Estimate=0.04, SE=0.12, $z$=0.39, $p$ =0.70).

### 3.2.5 Discussion

The results largely replicated the findings from Experiment 1. Neutral Fall received the lowest SI rate (54.5%), RFR the highest (62.5%), and the Concession Contour was in between numerically (57.4%). However, the differences were much smaller than in Experiment 1 such that only the comparison between Neutral Fall and RFR reached statistical significance. This compression may be due to the more mediated nature of the task: rather than judging one's own production - and by virtue of that most likely intention - the perception experiment required not only reasoning about the intention of someone else's choice of intonation but also how that might affect the hearer. The fact that the experiment was able to replicate the previous data is thus even more notable.

The next section turns to a discussion of what the results tell us about the meaning of the intonational contours involved.

## 4 Theoretical Implications

The main motivation of the two experiments presented above was to assess what effect the RFR has on SI calculation, testing the predictions made by previous accounts of the contour. Based on the finding that the RFR increased SI rate relative to a Neutral Fall, the experiments provide evidence against the accounts by Ward

& Hirschberg (1985), Wagner (2012), Wagner et al. (2013) and Westera (2019), compatible with Constant (2012), and in favor of those of Göbel (2019) and Göbel & Wagner (2023). Moreover, the result from the production rates that the RFR was almost exclusively produced in the STRONG condition provides additional support for the latter two accounts. However, we also saw (Figure 3) that there was substantial variation across items regarding the frequency at which the RFR was produced, which a satisfying theory should explain as well. While the items are too diverse to draw definitive conclusions, it is worth looking at potential patterns and seeing how they relate to accounts to inform future investigations.[2]

The first pattern of interest is that the RFR seemed to appear less frequently with adjectives than other syntactic categories: of the 60 items, 32 (53%) were adjectives, but were responsible for only 40% of produced RFRs; of the ten items with the most RFR occurrences, only two were adjectives; and of the items with one or zero occurrences, all but one were adjectives. A possible explanation for this pattern could come from Göbel & Wagner (2023), who argue that a question-answer context - as used in the experiments here - biases toward a use of the RFR that is concerned with a scale of informativity. In contrast, adjectives might be more likely to set up an evaluative scale, and therefore discourage the use of the RFR.

A second notable pattern is that the RFR occurred rarely with lexical scales that have some negative connotation, such as <*dirty*, *filthy*> or <*ugly*, *hideous*>. If confirmed, this observation may again receive an explanation on Göbel & Wagner's (2023) account: on the assumption that adjectives like *dirty* and *filthy* are on a measurement scale regarding cleanliness with other adjectives like *clean* and *pristine*, the stronger predicate *filthy* would actually be lower than *dirty* on the scale (see Solt 2015), which should make the RFR less acceptable.

Moving on from the RFR, another contour featured in this study is what we labeled the Concession Contour. As mentioned in the introduction, this contour resembles the Contradiction Contour, with an initial high tone followed by a pitch "valley" and a concluding rise. The most relevant question then is whether the Concession Contour is in fact just a different use of the Contradiction Contour that could receive a unified account. The feeling that the Contradiction Contour is usually more exaggerated could then be attributed to paralinguistic factors like emotional arousal.

To address this question, let's adopt the - to our knowledge, only available - formal account of the Contradiction Contour by Goodhue & Wagner (2018), according to which the contour presupposes contextual evidence for the complement of the prejacent. This account straightforwardly captures that the stereotypical use of the Contradiction Contour is when a previously asserted proposition *p* is contradicted

[2] As an additional caveat, it should be highlighted that low production rates don't necessarily entail low acceptability, insofar as production concerns a choice between possible options.

via $\neg p$. However, how would this account capture the use of the contour in our experimental conditions and its effect on SI rate? Given that the contour was used in replies to questions, one would have to assume that *?p* contributes contextual evidence for $\neg p$, which seems too strong. On the other hand, *?p* is usually only possible when $p$ - and $\neg p$ - is not known. The account could thus be adjusted to thinking of contextual evidence in terms of degrees (cf. Farkas & Roelofsen 2017's notion of credence levels). The increase in SI rate would thus follow from the contour presupposing that the negation of the prejacent - the strengthened interpretation - is in fact supported by some minimal amount of contextual evidence.

However, it is not clear how this account would capture the by-item variation we observed insofar as, especially in the SAME condition, the relation between the question and the response does not change. The question would then be why asking if *x* is hard would provide less contextual evidence for '*x* is not hard' than asking if *y* is permitted for '*y* is not permitted'. Given that the previous conception of the Contradiction Contour does not readily extend to our data, considering an account of the Concession Contour that is distinct from the Contradiction Contour seems therefore a justified option, which we will leave for future research.

As the last point, it is worth noting that the findings have important implications for the study of scalar diversity. As shown, intonational contours affect the likelihood of participants drawing an SI and the lexical material of a sentence affects how likely participants are to produce a certain contour. As a result, when comparing SI rates across different lexical scales using written stimuli, it is not easily discernible if differences are driven by the lexical scales themselves or mediated through lexical scales affecting rates of intonational contours. Future studies of scalar diversity should therefore control for effects of intonation.

## References

Baker, Rachel, Ryan Doran, Yaron McNabb, Meredith Larson & Gregory Ward. 2009. On the non-unified nature of scalar implicature: An empirical investigation. *International Review of Pragmatics* 1(2). 211–248. doi:10.1163/187730909x12538045489854.

Barr, Dale J, Roger Levy, Christoph Scheepers & Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3). 255–278. doi:10.1016/j.jml.2012.11.001.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. doi:10.18637/jss.v067.i01.

Beltrama, Andrea & Ming Xiang. 2013. Is 'good' better than 'excellent'? An experimental investigation on scalar implicatures and gradable adjectives. *Sinn und Bedeutung (SuB)* 17. 81–98.

Buccola, Brian & Daniel Goodhue. to appear. The effect of intonation on scalar and ignorance inferences. *Chicago Linguistic Society (CLS)* 59. https://ling.auf.net/lingbuzz/007464.

Büring, Daniel. 2003. On d-trees, beans, and b-accents. *Linguistics and Philosophy* 26. 511–545.

Chevallier, Coralie, Ira A Noveck, Tatjana Nazir, Lewis Bott, Valentina Lanzetti & Dan Sperber. 2008. Making disjunctions exclusive. *Quarterly Journal of Experimental Psychology* 61(11). 1741–1760.

Constant, Noah. 2012. English rise-fall-rise: a study in the semantics and pragmatics of intonation. *Linguistics and Philosophy* 35. 407–442.

Cummins, Chris & Hannah Rohde. 2015. Evoking context with contrastive stress: Effects on pragmatic enrichment. *Frontiers in Psychology* 6. 1779.

Doran, Ryan, Gregory Ward, Meredith Larson, Yaron McNabb & Rachel E. Baker. 2012. A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language* 88(1). 124–154.

Farkas, Donka F. & Floris Roelofsen. 2017. Division of labor in the interpretation of declaratives and interrogatives. *Journal of Semantics* 34. 237–289.

Goodhue, Daniel & Michael Wagner. 2018. Intonation, ‘yes’ and ‘no’. *Glossa: a journal of general linguistics* 3. 1–45.

Gotzner, Nicole. 2019. The role of focus intonation in implicature computation: A comparison with ‘only’ and ‘also’. *Natural Language Semantics* 27. 189–226.

Gotzner, Nicole, Stephanie Solt & Anton Benz. 2018. Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in Psychology* 9. 1659.

Grice, Herbert Paul. 1967. Logic and Conversation. In Paul Grice (ed.), *Studies in the Way of Words*, 41–58. Harvard University Press.

Göbel, Alexander. 2019. Additives pitching in: L*+H signals ordered Focus alternatives. *Semantics and Linguistic Theory (SALT)* 29. 279–299.

Göbel, Alexander & Michael Wagner. 2023. On a concessive reading of the rise-fall-rise contour: contextual and semantic factors. *Experiments in Linguistic Meaning (ELM)* 2. 83–94.

Hirotani, Masako. 2005. *Prosody and LF interpretation: Processing Japanese wh-questions*: University of Massachusetts, Amherst PhD dissertation.

Hirschberg, Julia Bell. 1985. *A theory of scalar implicature*: University of Pennsylvania PhD dissertation.

Höhle, Tilman N. 1992. Über Verum-Fokus im Deutschen. In Joachim Jacobs (ed.), *Informationsstruktur und grammatik*, 112–141. Opladen.

Horn, Laurence R. 1972. *On the semantic properties of logical operators in English*: UCLA PhD dissertation.

Hu, Jennifer, Roger Levy, Judith Degen & Sebastian Schuster. 2023. Expectations over Unspoken Alternatives Predict Pragmatic Inferences. *Association for Computational Linguistics* 11. 885–901. doi:10.1162/tacl_a_00579.

Hu, Jennifer, Roger Levy & Sebastian Schuster. 2022. Predicting scalar diversity with context-driven uncertainty over alternatives. *Workshop on Cognitive Modeling and Computational Linguistics* 68–74.

Jackendoff, Ray. 1972. *Semantic interpretation in generative grammar*. Cambridge, MA: MIT Press.

John M. Tomlinson, Jr, Nicole Gotzner & Lewis Bott. 2017. Intonation and pragmatic enrichment: How intonation constrains ad hoc scalar inferences. *Language and Speech* 60(2). 200–223. doi:10.1177/0023830917716101.

Levinson, Stephen C. 2000. *Presumptive Meanings*. MIT Press Ltd.

Liberman, Mark & Ivan Sag. 1974. Prosodic form and discourse function. *Chicago Linguistic Society (CLS)* 10. 416–427.

de Marneffe, Marie-Catherine & Judith Tonhauser. 2019. Inferring meaning from indirect answers to polar questions: The contribution of the rise-fall-rise contour. In Edgar Onea, Malte Zimmermann & Klaus von Heusinger (eds.), *Questions in discourse*, 132–163. Leiden: Brill.

Pankratz, Elizabeth & Bob van Tiel. 2021. The role of relevance for scalar diversity: a usage-based approach. *Language and Cognition* 13(4). 562–594. doi:10.1017/langcog.2021.13.

Roberts, Craige. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics* 5. 1–69. doi:https://doi.org/10.3765/sp.5.6. Earlier version appeared in OSU Working Papers in Linguistics 49 in 1996.

Ronai, Eszter & Ming Xiang. 2021. Exploring the connection between Question Under Discussion and scalar diversity. *Linguistic Society of America (LSA)* 6(1). 649–662. doi:10.3765/plsa.v6i1.5001.

Ronai, Eszter & Ming Xiang. 2022a. Quantifying semantic and pragmatic effects on scalar diversity. *Linguistic Society of America (LSA)* 7. 1–15.

Ronai, Eszter & Ming Xiang. 2022b. Three factors in explaining scalar diversity. *Sinn und Bedeutung (SuB)* 26. 716–733.

Simons, Mandy & Tessa Warren. 2018. A closer look at strengthened readings of scalars. *Quarterly Journal of Experimental Psychology* 71(1). 272–279. doi:10.1080/17470218.2017.1314516.

Solt, Stephanie. 2015. Measurement scales in natural language. *Language and Linguistics Compass* 9. 14–32.

Sun, Chao, Ye Tian & Richard Breheny. 2018. A link between local enrichment and scalar diversity. *Frontiers in Psychology* 9.

van Tiel, Bob, Emiel Van Miltenburg, Natalia Zevakhina & Bart Geurts. 2016. Scalar diversity. *Journal of Semantics* 33(1). 137–175. doi:10.1093/jos/ffu017.

Tomlinson, John Michael & Camilo R. Ronderos. 2021. Does intonation automatically strengthen scalar implicatures? *Semantics and Pragmatics* 14(4). 1–30. doi:10.3765/sp.14.4.

Wagner, Michael. 2012. Contrastive topics decomposed. *Semantics and Pragmatics* 5. 1–54.

Wagner, Michael, Elise McClay & Lauren Mak. 2013. Incomplete answers and the rise-fall-rise contour. *Semantics and Pragmatics of Dialogue (SemDial)* 17. 140–149.

Ward, Gregory & Julia Hirschberg. 1985. Implicating uncertainty: the pragmatics of fall-rise intonation. *Language* 61. 747–776.

Westera, Matthijs. 2019. Rise-fall-rise as a marker of secondary QUDs. In Daniel Gutzmann & Katharina Turgay (eds.), *Secondary content: the linguistics of side issues*, 376–404. Leiden: Brill.

Westera, Matthijs & Gemma Boleda. 2020. A closer look at scalar diversity using contextualized semantic similarity. *Sinn und Bedeutung (SuB)* 24. 439–454.

Alexander Göbel
Program in Linguistics
1-S-19 Green Hall
Princeton, New Jersey 08544
alexander.gobel@princeton.edu

Eszter Ronai
2016 Sheridan Rd
Room 205
Evanston, IL 60208
ronai@northwestern.edu