Processing Scalar Inferences in Face-Threatening Contexts

Jean-François Bonnefon (bonnefon@univ-tlse2.fr)
CNRS and Université de Toulouse
Toulouse, France

Wim De Neys (deneys@univ-tlse2.fr)
CNRS and Université de Toulouse
Toulouse, France

Aidan Feeney (a.feeney@qub.ac.uk)
Queen's University Belfast
Belfast, Northern Ireland

Abstract

Depending on politeness considerations, the quantifier 'some' can receive a broad interpretation (some and possibly all) or a narrow interpretation (some but not all). Face-threatening statements such as 'some people hated your speech' encourage the broad interpretation that everyone hated the speech. Because previous research showed that broad interpretations are normally faster and easier, politeness should be easy to process, since it would encourage what is normally the easier interpretation of the statement. Using response time measures and a cognitive load manipulation, this research shows that just the opposite is true: Face threatening contexts encourage the broad interpretation of 'some' while making it longer and more difficult to reach. This result raises difficulties for current cognitive theories of pragmatic inferences.

Keywords: Scalar inference; politeness; processing; response time; cognitive load.

The Inference from 'Some'

Experimental pragmatics engages in an experimental investigation of the mental processes involved in inferring what people mean from sentences, contexts, and the implicit principles that govern the use of sentences in context (Noveck & Reboul, 2008). The drosophila of experimental pragmatics is the contextual inference from 'some' to *some but not all.* Although 'some' is semantically consistent with the broad interpretation *some and possibly all*, it is commonly given the narrow interpretation *some but not all.* For example, most adult speakers of English would assume the assertion of (1-a) to convey that the speaker believes (1-b):

(1) a. Some of the guests brought wine.
b. Not all of the guests brought wine.

The inference from (1-a) to (1-b) is a *scalar* inference (Horn, 1984), stemming from the ordered informativeness scale $<$ some, all $>$. From a Gricean perspective (Grice, 1989), if a cooperative speaker were in a position to assert (1-b), he or she would do so, and would not use the less informative (yet logically consistent) wording (1-a).

A key theoretical issue within experimental pragmatics is to explain why and how people adopt the narrow interpretation of 'some' in some contexts but not in others. Bonnefon, Feeney, and Villejoubert (2009) in particular showed that broad interpretations were more likely in face-threatening contexts, for reasons of politeness; but they did not elucidate *how* individuals adopted this interpretation. This is the question we will address.

Face-Threatening Contexts

Face is a sense of public self-esteem that all of us project, and are motivated to support, in social interactions. Many actions, called face-threatening acts, can induce a loss of face (e.g., apologizing, or criticizing). Performing such an action often requires the use of a politeness strategy that mitigates the face threat (Brown & Levinson, 1978/1987).

One of these strategies is to hedge by means of a scalar term. For example, instead of bluntly asserting (2-a), speakers may politely hedge as in (2-b):

(2) a. You are wrong.
b. You are possibly wrong.

When they hear (2-b), individuals tend to interpret 'possibly' as denoting a high probability, because they do not construe this term as a genuine expression of uncertainty, but rather as a polite way to preserve the face of the listener (Bonnefon & Villejoubert, 2006; Pighin & Bonnefon, in press). In a sense, people usually adopt a narrow interpretation of 'possibly' that eliminates high probabilities, but they switch to a broad interpretation (that includes high probabilities) in face-threatening contexts.

Comparably, instead of bluntly asserting (3-a), speakers may politely hedge as in (3-b):

(3) a. All of the guests hated the meal you cooked.
b. Some of the guests hated the meal you cooked.

When they hear (3-b), individuals tend to disregard the scalar inference attached to 'some' and to reach a broad interpretation. That is, when X in 'some X-ed' threatens the face of the listener, individuals find it likely that the speaker might have meant that all X-ed, but used the scalar term nonetheless, out of politeness (Bonnefon et al., 2009). As we will now explain, the current article aims at elucidating the time and effort it takes to reach that interpretation, and thus at elucidating whether politeness is easy or hard to process.

Processing Scalar Inferences

Experimental pragmatics has devoted considerable attention to the cognitive processes underlying broad and narrow interpretations of 'some'. This experimental work stemmed from the controversy between generalized and particularized approaches to scalar inferences. Generalized approaches (e.g., Levinson, 2000) claim that scalar inferences are endorsed by default, but then canceled in some contexts: Narrow interpretations should thus be faster and easier than broad interpretations. Particularized approaches, not limited to but often identified with Relevance Theory (Sperber & Wilson, 1986/1995), claim that scalar inferences are not derived by default, but rather triggered in some contexts and not in others: Broad interpretations should thus be faster and easier than narrow interpretations.

The extant evidence broadly supports the particularized view. The narrow interpretation appears to take longer than the broad interpretation (Bott & Noveck, 2004; Noveck & Posada, 2003), the narrow interpretation is less frequent when participants must respond quickly (Bott & Noveck, 2004), and it is less frequent when participants must carry out a secondary task (De Neys & Schaeken, 2007). Finally, the processing time of 'some' decreases in contexts that make the narrow interpretation inappropriate (Breheny, Katsos, & Williams, 2006). Prior data thus suggest that broad interpretations of 'some' are faster and easier than narrow interpretations.

The question remains, though, of whether this pattern of results will hold in face-threatening contexts. As summarized in the previous section, we already know that face-threatening contexts encourage broad interpretations. If we show that broad interpretations remain faster and easier in face-threatening contexts (as they are in other contexts), then we will be in a position to argue that politeness is easy to process: Politeness considerations would quickly direct people to the easy interpretation of 'some'. The possibility arises, however, that politeness is in fact *hard* to process. Stephan, Liberman, and Trope (2010), in particular, provided data suggesting that politeness was associated with abstract levels of cognitive construal, which are assumed to tap into effortful cognitive processes. If really politeness is hard to process, then people should slowly and effortfully reach the broad interpretation of 'some' in face-threatening contexts, in contrast to standard results.

The question of whether face-threatening contexts make broad interpretations easier or harder is an important one for experimental pragmatics. The particularized approach to scalar inferences (currently supported by most available data) assumes that any context that prompts a broad interpretation results in easier processing. Showing that face-threatening contexts encourage broad interpretations whilst making them harder would require to revisit this basic assumption.

In the rest of this article, we provide experimental data investigating whether face-threatening contexts make broad interpretation easy and fast, or difficult and slow. We contrast face-threatening statements (e.g., some people hated your speech) with their control version (e.g., some people loved your speech). We record response times associated with broad and narrow interpretations of some, for both types of statement. Previous research suggests that broad interpretations will be faster for 'love' statements. If politeness is easy to process, this result will extend to 'hate' statements; but if politeness is hard to process, this result will be reversed for 'hate' statements.

In parallel, we record the percentage of broad interpretations of 'love' and 'hate' statements reached under cognitive load, to that reached without cognitive load. Previous research suggests that cognitive load will increase the proportion of broad interpretations for 'love' statements. If politeness is easy to process, this result will extend to 'hate' statements; but if politeness is hard to process, this result will be reversed for 'hate' statements.

Method

A total of 356 first-year psychology students of the University of Leuven participated in return for course credit.

Scalar interpretation task

Participants were presented with two problems (the 'trip' and 'speech' scenarios), in random order for each participant. Half of the participants read a face-boost version of the two scenarios whereas the other half read a face-threat version of the two scenarios. Below is an example of the face-boost version ('trip' scenario), translated from Dutch :

> *Imagine you organized a group trip. You are discussing the trip with Alice, who was in the group. There were 6 other people who went on this trip. You are considering whether to recommend the trip to some friends.*
>
> Hearing this, Alice tells you that **'Some people loved the way the trip was organised.'**
>
> Given what Alice tells you, do you think that it is possible that everybody loved the way the trip was organized?
>
> 1. Yes
> 2. No

A 'Yes' response indicates a broad interpretation, and a 'No' response indicates a narrow interpretation. In the face-threat condition, the word 'loved' was replaced with the word 'hated'.

In its face-boost version, the The 'speech' scenario read:

> *Imagine you gave a speech at a small political rally. You are discussing your speech with Denise, who was in the audience. There were 6 other people in the audience. You are considering whether to give this same speech to another audience.*
>
> Hearing this, Denise tells you that **'Some people loved your speech.'**
>
> Given what Denise tells you, do you think that it is possible that everybody loved your speech?

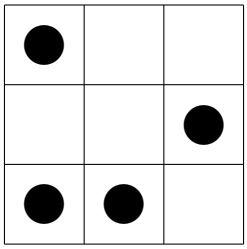

Figure 1: Example of dot pattern in the load group.

1. Yes
2. No

A 'Yes' response indicates a broad interpretation, and a 'No' response indicates a narrow interpretation. In the face-threat condition, the word 'loved' was replaced with the word 'hated'.

The problems were presented on a computer screen. Participants first read the background information (text in italics in the example above). When they were finished, they pressed the ENTER-key and the remaining part of the scenario was presented. Response time was measured starting at the moment participants pressed the key, and thus does not include the time participants spent reading the background information.

Dot memory task

Cognitive load was manipulated by asking half of the participants in the face-threat and face-boost conditions to memorize a dot pattern while making their scenario judgments. A 3×3 matrix filled with a complex 4-dot pattern (see Figure 1) was briefly presented after participants had read the background information of the scenario. Storage of these complex dot patterns has been shown to tap executive resources (De Neys, 2006; Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001). Participants memorized the pattern and were asked to reproduce it after they had entered their response to the scenario question.

Procedure

Participants were tested in groups of 18 to 31. Approximately half of the participants were assigned to the face-threat and face-boost group. Within each group half of the participants were assigned to the load and control condition. In the load condition the dot pattern was presented for 850 ms after participants indicated that they were finished reading the background information by pressing the ENTER-key. Next, participants were presented with the remaining part of the scenario and entered their response. No specific instructions were given to participants as to how long they should think about their response. After participants had entered their response, an empty matrix was presented and participants had to reproduce the dot pattern. They received feedback on whether the pattern had been reproduced correctly and were reminded that they had to remember the complete pattern correctly. The load procedure was clarified with two practice items. Instructions stressed that it was crucial that the dot patterns were reproduced correctly in the upcoming task.

Results

Dot memory task

The dot memory task was properly performed, with an average of 3.5 dots correctly localized.

Percentage of narrow interpretations

Figure 2 displays the percentage of scalar inferences (as measured by a narrow interpretation of 'some') as a function of cognitive load, in the face-boost context as well as in the face-threat context. An analysis of variance[1] failed to detect any main effect of load, $F(1,352) < 1$, $p > .25$ (unless otherwise mentioned, p values are 1-tailed), but detected a main effect of context, $F(1,352) = 10.4$, $p < .001$, reflecting the fact that narrow interpretations were more frequent in the face-boost context, overall, than in the face-threat context.

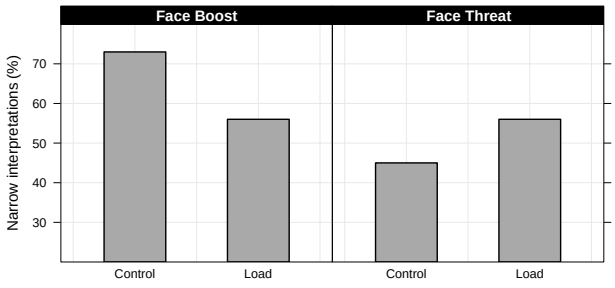

Figure 2: Percentage of scalar inferences, as measured by a narrow interpretation ('No' responses) of 'some'. Cognitive load decreases the proportion of scalar inferences in the face-boost context, but it has the opposite effect in the face-threat context.

Critically, this main effect was qualified by a significant interaction between load and context, $F(1,352) = 9.9$, $p < .001$. As already apparent from the visual inspection of Figure 2, planned comparisons revealed that load had opposite effects in the face-boost and face-threat contexts. In the face-boost context, load significantly decreased the frequency of narrow interpretations (from 73% down to 56%), $F(1,352) = 7.0$, $p < .01$. In contrast, in the face-threat context, load *increased* the frequency of narrow interpretations (from 45% up to 56%), $F(1,352) = 3.2$, $p < .05$.[2]

[1] Because each participant saw two scenarios, the dependent variable in this analysis is the average individual number of narrow interpretations. No differences were detected between the responses to the two scenarios.

[2] The rate of narrow interpretations in the load condition remains significantly above chance level, $t = 1.93$, $p = .05$ (two-tailed), ruling out the possibility that participants in the load condition responded randomly under cognitive load.

Response times in the control condition

Because caution is required in interpreting response time data under load, we focus our analyses on the response time data in the control condition.[3]

Table 1: Average response time (in seconds) for broad and narrow interpretations, in the face-boost and face-threat contexts. Standard deviations are included in parentheses.

	Broad Responses	Narrow Responses
Face-Boost	11.1(5.9)	10.1(4.1)
Face-Threat	13.9(6.8)	10.2(4.2)

Table 1 displays the time taken to reach narrow or broad interpretations of 'some', in the face-boost and face-threat contexts. An analysis of variance detected a main effect of the type of response, $F(1,217) = 10.1$, $p < .01$, reflecting the fact that broad responses took overall longer than narrow responses, and a main effect of context, $F(1,217) = 3.8$, $p < .05$, reflecting the fact that participants in the face-threat context took longer to respond. These two main effects, however, were qualified by an interaction effect, $F(1,217) = 3.1$, $p < .05$, which could readily be anticipated from Table 1. The difference in response time between the broad and the narrow interpretations was only significant in the face-threat context. In the face-boost condition, broad interpretations of 'some' took about the same time as narrow interpretations, $t(43.2) = 1.1$, $p = .30$ (corrected for unequal variances). In the face-threat condition, however, broad interpretations of 'some' took much *longer* than narrow interpretations, and this difference achieved significance, $t(109.4) = 3.5$, $p < .001$ (corrected for unequal variances).

For an alternate view of the response time data, Figure 3 displays the proportion of narrow interpretations in the two contexts, as a function of response speed (equal-width intervals of 7 s). In the face-boost condition, the proportion of narrow responses was not detectably different as a function of response speed, $F(2,93) < 1$, $p = .68$. Results in the face-threat condition, however, clearly suggest that the broad interpretation was slowly acquired, with narrow interpretations being less and less likely for slower responses, $F(2,93) = 5.7$, $p < .01$.

Discussion

The quantifier 'some' is commonly given the narrow interpretation *some but not all* rather than the broad interpretation *some and possibly all,* but this tendency varies across

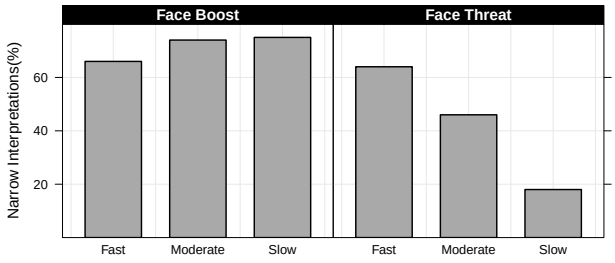

Figure 3: Percentage of narrow interpretations ('No' responses) as a function of response speed, in the no load condition. In the face-boost condition, narrow interpretations dominate however fast or slow the response. In the face-threat context, narrow interpretations dominate among fast responders, but broad interpretations are gradually preferred as responses get slower.

contexts. In particular, it is inverted in face-threatening contexts, wherein the broad interpretation of 'some' threatens the positive self-image of the listener. Up until now, research in experimental pragmatics always found that the broad interpretation of 'some' reflected faster and less demanding cognitive processing than its narrow interpretation. This article aimed at investigating whether this result would reverse in face-threatening contexts, in line with the hypothesis that politeness is hard to process. We asked 356 participants to interpret 'some' statements in a face-threatening context, or in a control, face-boosting context. In both contexts, we recorded the response time associated with the narrow and broad interpretations of 'some'. Furthermore, and still in both contexts, we assessed the effect of cognitive load on the interpretation of 'some'.

Results in the face-boost condition were consistent with previous findings. Cognitive load did reduce the frequency of narrow interpretations (De Neys & Schaeken, 2007), although narrow interpretations were not reliably associated with a longer response time (Bott & Noveck, 2004; Noveck & Posada, 2003). It is possible that our response time measure was not sensitive enough to capture an effect that previous findings suggest to be small, and not always reproducible (Feeney, Scrafton, Duckworth, & Handley, 2004; Grodner, Klein, Carbary, & Tanenhaus, 2010).

In the face-threat context, though, cognitive load *increased* the frequency of narrow interpretations, and *broad* interpretations took longer. Taken together, these findings suggest that broad interpretations are no longer quick and easy to reach when the context is face-threatening, but that they rather require slow and effortful processing. We draw this conclusion with due caution, for we only used two scenarios in this experiment, which makes it difficult to control for noisy response time measures. In the rest of this article, we examine the theoretical implications of this finding for the particularized approach to scalar inferences, and more specifically to

[3]The pattern of response time in the load condition was similar to that we report for the control condition, it is left out only because its interpretation would require multiple qualifications. In the face-boost condition, broad interpretations of 'some' took about the same time as narrow interpretations (12.4 s vs. 12.1 s); and in the face-threat condition, broad interpretations of 'some' took longer than narrow interpretations (13.3 s vs. 11.6 s).

one of its best-established models, Relevance Theory.

Theoretical implications

Previous work suggested that contexts which increased the frequency of broad interpretations also made such interpretations easier (Breheny et al., 2006). Clearly, face-threatening contexts do not work that way. On the contrary, they appear to encourage broad interpretations whilst making these very interpretations more difficult. Politeness, thus, is not processed in a fast and effortless way; it rather appears to add a layer of complexity to the usual processes involved in the interpretation of 'some'. The question is whether these processes can be captured by current cognitive models of scalar inference.

Relevance theory (Sperber & Wilson, 1986/1995) is among the best-established of these cognitive models. Relevance theory assumes that people settle on the optimally relevant interpretation of a statement, which requires that (a) the statement achieves cognitive effects that are large enough to make it worth processing; and (b) these effects could not have been achieved by another statement, that would be easier to process and compatible with the communicator's abilities and preferences.

Relevance theory can explain why a broad interpretation of some is optimally relevant in a face-threatening context. It is a plausible assumption that the cognitive effect of reaching a face-threatening interpretation (e.g., everybody hated your speech) is worth what our data suggest to be a considerable processing effort, and although this cognitive effect might be achieved for less effort, by bluntly stating the face-threatening fact, this would go against a plausible preference of the speaker, that of being nice.

One issue remains, though. Relevance theory assumes that people adopt a path of least effort to interpretation, and that they stop at the first optimally relevant interpretation. Our data show that in face-threatening contexts, the narrow interpretation of 'some' requires less time and effort than its broad interpretation. That the broad interpretation be optimally relevant is therefore not enough to explain why people adopt it, as it is also necessary to explain why they do not stop first at the narrow interpretation. Only two theoretical possibilities arise: (a) the narrow interpretation is not optimally relevant in face-threatening contexts; or (b) the narrow interpretation is optimally relevant, but people still continue searching for another interpretation, and eventually settle on the broad interpretation.

Both options come with their own set of difficulties. The first option would require us to explain why, processing costs being equal, the cognitive effects of a narrow interpretation would be weaker in face-threatening contexts, as compared to other contexts. The second option is consistent with the view that politeness contexts prompt people to look for covert, indirect meaning (Demeure, Bonnefon, & Raufaste, 2008), but it would require to let go of a fundamental assumption of Relevance theory. It would indeed require to accept that listeners, when cued that politeness is required by the situation, do not stop at the first equilibrium between effect and effort, but tentatively engage in a second push towards another equilibrium.

We can attempt to capture the processing consequences of this last option by fitting a simple mathematical model to our data. We start by assuming that a given proportion P_1 of participants is likely to make the first move to a narrow interpretation of 'some *x*-ed', whether *X* is a face threat or a face boost. Let us now assume that when *X* is a face threat, some proportion P_2 of participants not only make the first move to the narrow interpretation, but also a second move to the polite broad interpretation. Let us finally model concurrent load with two parameters $\alpha < 1$ and $\beta < 1$, such that only $\alpha \times P_1$ of the participants are able to overcome load when making the first move, and only $\beta \times P_2$ of the participants are able to overcome load when making the second move. Fitting this simple model to our data, we arrive at $P_1 = .73$, $P_2 = .28$, $\alpha = .77$, and $\beta = 0$.

According to these parameter values, our data would suggest that arriving at the first equilibrium associated with the narrow interpretation is relatively easy and natural: Three quarters of participants would make it under control conditions, and three quarters of these would continue to do so under load (note that the relative easiness of this move is consistent with our response time data). However, arriving at the second politeness cued equilibrium is much more effortful. Only a quarter of participants would reach this second equilibrium when there is no load, and none would be able to reach it under load.

Our data cannot, on their own, help decide whether Relevance Theory should be modified in line with this model, or whether it should be modified so as to argue that narrow interpretations are not optimally relevant in face-threatening contexts. Rather, they put the theory at an exciting crossroad, where each route comes with its set of theoretical obstacles. This is, one might argue, a success for experimental pragmatics. Not only can linguistic theories be tested through experimental data, but experimental data can also sometimes force hard choices on a well-tested theory.

Acknowledgments

Jean-François Bonnefon and Wim De Neys, CLLE (CNRS, UTM, EPHE), Toulouse, France. Aidan Feeney, School of Psychology, Queen's University Belfast, Northern Ireland. Address correspondence to Jean-François Bonnefon, 'Cognition, Langues, Langage et Ergonomie,' Maison de la recherche, 5 allées A. Machado, 31058 Toulouse Cedex 9, France. E-mail: bonnefon@univ-tlse2.fr. We thank Gaëlle Villejoubert for useful discussions about this work. Data were collected while Wim De Neys was affiliated to the Lab Experimentele Psychologie, K.U. Leuven, Leuven, Belgium. This research was supported by grant ANR-07-JCJC-0065-01 from the Agence Nationale de la Recherche, an FWO (Fonds Wetenschappelijk Onderzoek-Vlaanderen) research grant, and by grant PN 08.010 (awarded by the British Council) and 18223VF (awarded by the Ministère des Af-

faires Étrangères et Européennes) under the Alliance Franco-British Programme 2008.

References

Bonnefon, J. F., Feeney, A., & Villejoubert, G. (2009). When some is actually all: Scalar inferences in face-threatening contexts. *Cognition*, *112*, 249–258.

Bonnefon, J. F., & Villejoubert, G. (2006). Tactful or doubtful? Expectations of politeness explain the severity bias in the interpretation of probability phrases. *Psychological Science*, *17*, 747–751.

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Cognition*, *51*, 437–457.

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, *100*, 434–463.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press. (Original work published 1978)

De Neys, W. (2006). Dual processing in reasoning: Two systems but one reasoner. *Psychological Science*, *17*, 428–433.

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load –Dual task impact on scalar implicatures. *Experimental Psychology*, *54*, 128–133.

Demeure, V., Bonnefon, J. F., & Raufaste, E. (2008). Utilitarian relevance and face management in the interpretation of ambiguous question/request statements. *Memory and Cognition*, *36*, 873–881.

Feeney, A., Scrafton, S., Duckworth, A., & Handley, S. J. (2004). The story of *some:* Everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology*, *58*, 121–132.

Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: MIT Press.

Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). "Some", and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, *116*, 42–55.

Horn, L. R. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, form, and use in context* (pp. 11–42). Washington, DC: Georgetown University Press.

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

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, *130*, 621–640.

Noveck, I. A., & Posada, A. (2003). Characterising the time course of an implicature. *Brain and Language*, *85*, 203–210.

Noveck, I. A., & Reboul, A. (2008). Experimental pragmatics: a Gricean turn in the study of language. *Trends in Cognitive Sciences*, *11*, 425–431.

Pighin, S., & Bonnefon, J. F. (in press). Facework and uncertain reasoning in health communication. *Patient Education and Counseling*.

Sperber, D., & Wilson, D. (1995). *Relevance, communication and cognition*. Oxford: Blackwell. (Original work published 1986)

Stephan, E., Liberman, N., & Trope, Y. (2010). Politeness and psychological distance: A construal level perspective. *Journal of Personality and Social Psychology*, *98*, 268–280.