# 14 The why and how of experimental pragmatics: the case of 'scalar inferences'

*Ira Noveck and Dan Sperber*

Although a few pioneers in psycholinguistics had taken an experimental approach to various pragmatic issues for more than twenty years, it is only in the past few years that investigators have begun using experimental methods to test pragmatic hypotheses (see Noveck and Sperber 2004). We see this emergence of a proper experimental pragmatics as an important advance, with great potential for further development. In this chapter we want to illustrate what can be done with experimental approaches to pragmatic issues by considering one example, the case of so-called 'scalar inferences', where the experimental method has helped sharpen a theoretical debate and provided uniquely relevant evidence. We will focus on work done by the first author and his collaborators, or work closely related to theirs, but other authors have also made important contributions to the topic (e.g. Papafragou and Musolino 2003; Guasti, Chiercha, Crain, Foppolo, Gualmini and Meroni 2005; De Neys and Schaeken 2007).

## 14.1 Methodological background: the limits of pragmatic intuitions as evidence

Theoretical work in pragmatics relies heavily – often exclusively – on pragmatic intuitions. These are rarely complemented with observational data of a kind more common in sociologically oriented pragmatics. Use of statistical data from corpuses or experiments is even less common. This is partly a result of the fact that most theoretical pragmaticists are trained in departments of linguistics, where linguistic intuitions are quite often the only kind of data considered. Optimally, of course, one would want pragmaticists to use whatever kind of data might significantly confirm or disconfirm hypotheses. Moreover, a sensible desire for methodological pluralism is not the only reason to diversify the types of evidence used in pragmatics. There are also principled limits to the use of pragmatic intuitions.

It makes sense (although it is not entirely uncontroversial) to judge a semantic description by its ability to account for semantic intuitions. Of course the use of semantic intuitions, and of linguistic intuitions generally, raises methodological

problems, and calls for methodological caution. For instance, a linguist's intuitions may be biased by prior theoretical commitments. One can also mistake what are in fact pragmatic intuitions for semantic ones (as ordinary language philosophers systematically did, according to Grice). Still, there are good reasons why semantic intuitions are so central to semantics. Semantic intuitions are not just *about* semantic facts; they are themselves semantic facts. For instance, the intuition that sentence (1) entails (2) is not *about* some semantic property that this sentence would have anyhow, whether or not it was accessible to speakers' intuitions.

(1) John knows that it is raining.

(2) It is raining.

Rather, for (1) to have the meaning it has *is* (among other things) for it to be intuitively understood as entailing (2). A semantic analysis of linguistic expressions that accounts for all the speaker–hearer's semantic intuitions about these expressions may not be the best possible one, but it is descriptively adequate (in Chomsky's sense). By contrast, an explanatorily adequate description of the semantics of a given language involves hypotheses about the capacities involved in the acquisition of this semantics, and here observational and experimental evidence should be of relevance.

The use of pragmatic intuitions raises the same methodological problems as the use of semantic intuitions, and more besides. It is a mistake to believe that the type of pragmatic intuitions generally used in pragmatics are data of the same kind as the semantic intuitions used in semantics. Genuine pragmatic intuitions are the intuitions hearers have about the intended meaning of utterances addressed to them. However, the pragmatic intuitions appealed to in theoretical pragmatics are not normally about actual utterances addressed to readers of a pragmatics article, but about hypothetical cases involving imaginary or generic interlocutors. Pragmatic intuitions about hypothetical utterances have proved useful in a variety of ways, but it is important to keep in mind that they are not intuitions about how an utterance is interpreted, but about how an utterance *would be* interpreted if it were produced in a specific situation by a speaker addressing an actual hearer, with referring expressions being assigned actual referents, and so on. These intuitions are educated guesses – and no doubt generally good ones – about hypothetical pragmatic facts, but they are not themselves pragmatic facts, and they may well be in error. That is, we may be wrong about how we would in fact interpret a given utterance in a given context.

Apart from helping to compensate for the inherent limitations of pragmatic intuitions, an experimental approach can provide crucial evidence that helps to choose between alternative theories which may assign the same interpretive content to utterances, but have different implications for the

cognitive mechanisms used in arriving at these interpretations. To make a worthwhile contribution, of course, experimentalists must conform to fairly strict methodological criteria and measure exactly what they are aiming to measure – typically the effect of one 'independent' variable on another 'dependent' variable without other uncontrolled variables affecting the results. We will show how this plays out in the study of 'scalar inferences'.

## 14.2 Theoretical background: scalar implicatures as Generalised Conversational Implicatures (GCIs)

The experiments we will present are relevant to the study of so-called 'scalar implicatures'. Here we briefly remind readers of the main features of the Gricean and neo-Gricean accounts of these, and focus on the claim that scalar implicatures are Generalised Conversational Implicatures, or GCIs. Scalar implicatures are illustrated by examples such as (3a), which is said to implicate (3c), or (4a), said to implicate (4c):

(3) a. It is possible that Hillary will win.
b. It is certain that Hillary will win.
c. It is not certain that Hillary will win.

(4) a. Some of the guests have arrived.
b. All of the guests have arrived.
c. Not all of the guests have arrived.

Proposition (3b) is more informative than (3a), which it entails. If the more informative proposition would make a greater contribution to the common purpose of the conversation, then a speaker obeying Grice's first Maxim of Quantity ('Make your contribution as informative as is required') would be expected to express it unless she were unable to do so without violating the Supermaxim of Quality ('Try to make your contribution one that is true'). Thus, on a Gricean account, a speaker uttering (3a) typically implicates (3c) (i.e. the negation of (3b)). For the same reasons, a speaker uttering (4a) typically implicates (4c) (i.e. the negation of (4b)).

These implicatures are described as 'scalar' because, according to an account developed by neo-Griceans, and in particular by Laurence Horn (1972), their derivation draws on pre-existing linguistic scales consisting of a set of alternate terms or expressions ranked by order of informativeness: <*possible, certain*> and <*some, all*> are examples of such scales. When a less informative term on a scale is used in a way that appears not to satisfy the first Maxim of Quantity, the speaker can be taken to implicate that the proposition that would have been expressed by use of a stronger term is false. This account of the type of implicatures carried by (3a) or (4a) extends to a wide variety of cases, and has some intuitive appeal. However, it should not be seen as obviously correct or as

having no alternative. In particular, its implications for processing are less attractive. According to this account, the inference from the utterance to its scalar implicature goes through a consideration not only of what the speaker said and the context, but also of what the speaker might have said but did not. It is this type of effort-demanding inference that makes the Gricean account of implicature derivation seem implausible from a cognitive and developmental point of view.

Levinson draws on another idea of Grice's, that of Generalised Conversational Implicatures, to propose an account that might offer a solution to the problem posed by the derivational complexity of scalar implicatures. Grice noted that some implicatures are generally valid (from a pragmatic rather than a logical point of view, of course) and could therefore be inferred without taking the context into account, except in the small number of cases where the context happens to make them invalid. Grice contrasted these Generalised Conversational Implicatures with Particularised Conversational Implicatures, which are valid only in certain contexts. In his book *Presumptive Meanings: The Theory of Generalized Conversational Implicature* (Levinson 2000), Levinson elaborates Grice's original and somewhat vague notion. For Levinson, GCIs are *default inferences*, that is, inferences which are automatically generated but can be cancelled in certain contexts. Levinson treats scalar implicatures as paradigm cases of GCIs (whereas Grice's own examples of GCIs do not include scalar implicatures). This proposal has the advantage of making the derivation of these implicatures a relatively simple one-step process, which needs no access either to contextual premises or to the full Gricean rationale for their derivation.

Levinson's own rationale for GCIs so conceived has to do with the optimisation of processing. The existence of GCIs speeds up the communication process, which Levinson argues is slowed down by the need for phonetic articulation: some unencoded aspects of the speaker's meaning can be inferred from metalinguistic properties of the utterance such as the choice of a given word from a set of closely related alternatives. For instance, the speaker's choice of 'some' rather than the stronger 'all' in (4a) ('Some of the guests have arrived') justifies the inference that (4c) is part of her meaning. These are non-demonstrative inferences, of course. There are cases where they would be invalid. For instance, if it is clear in the context that the speaker of (4a) has only partial information about the arrival of the guests, then (4c) would not be part of her meaning. Still, given that GCIs are valid in most contexts (or so it is assumed), the overall increase in the speed of communication brought about by their automaticity is not compromised by the rare cases where they have to be cancelled for contextual reasons.

The theory of scalar implicatures as default GCIs makes four claims:

(a) These inferences are made by default, irrespective of the context, and cancelled when required by the context.

(b) The fact that these inferences are made by default adds to the speed and efficiency of communication.
(c) These inferences contribute to utterance interpretation at the level of implicatures, rather than as enrichments of its explicit content (in Grice's terms, *what is said*, or in relevance theory's terms, its *explicatures*).
(d) These inferences are scalar: they exploit pre-existing scales such as <*some*, *all*>, <*or*, *and*>, <*possible*, *necessary*>.

We doubt all four claims. The bulk of this chapter will be devoted to explaining how experimental evidence has cast strong doubts on claim (a). First, however, we briefly present an argument which also casts doubt on (b), and outline the relevance-theoretic approach, which is in contradiction with all four claims.

This idea that default implicatures or GCIs would increase the speed and efficiency of communication may seem sensible and capable of lending support to the whole theory. However, it raises the following empirical issue. If GCIs had to be cancelled too often, their cost would offset the benefit of deriving them by default. Suppose, for instance, that a certain type of GCI had to be cancelled a third of the time. The total cost of using such a GCI would be the cost of deriving it by default in all cases, plus the cost of cancelling it in a third of cases. This would have to be compared with the cost of deriving the implicature as a 'particularised conversational implicature' – that is, in a context-sensitive and therefore costlier way – in two-thirds of the cases, but without the cost of default derivation followed by cancellation in the other third of cases. It is not clear that, given such frequencies, the proposed rationale for GCIs in terms of economy would make much sense.

To show that this kind of calculus is not unrealistic, consider the example of 'P or Q' and its alleged GCI *not (P and Q)*. We are not aware of any statistical data on the frequency of exclusive uses of 'or', and we share the common intuition that often, when people utter a sentence of the form 'P or Q' they can be taken to exclude the possibility that both *P* and *Q* are true. However, it does not follow that this is part of their meaning. In most cases, the fact that *P and Q* is excluded follows from real world knowledge and not from the interpretation of 'or', as illustrated in (5)–(7):

(5) He is a bachelor or he is divorced.

(6) Jane is in Paris or in Madrid.

(7) Bill will arrive Monday or Tuesday.

If 'P or Q' implicates by default that *not (P and Q)*, then in cases such as (5)–(7) where the two disjuncts cannot both be true for common-sense reasons, people will compute a GCI that makes them understand the speaker as redundantly implicating what is already part of the common ground, and this is surely a cost without an associated benefit. Moreover, if we are careful to exclude cases

where the mutual exclusivity of the disjuncts is self-evident and need not be communicated, and look only at cases such as (8)–(10), where neither the inclusive nor the exclusive interpretation is ruled out a priori, it is not at all obvious that the exclusive interpretation of 'or' is dominant:

(8) She wears sunglasses or a cap.

(9) Our employees speak French or Spanish.

(10) Bill will sing or play the piano.

We have no hard statistical data to present, but it seems less than obvious that a disposition to understand utterances of the form 'P or Q' by default as implicating *not (P and Q)* would increase the speed or efficiency of communication. More generally, the effect that GCIs would have on the efficiency of communication should be investigated rather than assumed.

## 14.3 Relevance theory's approach

We will assume that the basic tenets of relevance theory are familiar (Sperber and Wilson 1995; see also Wilson and Sperber 2004 for a recent restatement), and focus on how it applies to what neo-Griceans call 'scalar implicatures'. Two basic ideas play a crucial role here:

(a) Linguistic expressions serve not to *encode* the speaker's meaning but to *indicate* it. The speaker's meaning is inferred from the linguistic meaning of the words and expressions used, together with the context.

(b) The speaker's explicit and implicit meaning (her explicatures and implicatures) are inferred not sequentially but in parallel. The final overall interpretation of an utterance results from mutual adjustment of implicatures and explicatures guided by expectations of relevance

Here is a simple illustration of these two points:

(11) HENRY: Do you want to go on working, or shall we go to the cinema?
JANE: I'm tired. Let's go to the cinema.

Jane's description of herself as 'tired' achieves relevance by explaining why she is accepting Henry's suggestion. It must therefore be understood as conveying not simply that she is tired, but that she is too tired to go on working, while at the same time not too tired to go to the cinema. The word 'tired' is used to indicate an *ad hoc* concept TIRED*, with an extension narrower than that of the linguistically encoded concept TIRED. Whereas TIRED extends from a minimal level of tiredness to complete exhaustion, TIRED* extends only over those levels of tiredness that explain why Jane would rather go to the cinema than work. Henry correctly understands Jane's explicature to be (12) and her implicature to be (13), the result being an optimally relevant interpretation:

(12) I am TIRED*

(13) The reason why I would rather go to the cinema than work is that I am TIRED*

Note that the explicature in (12), and in particular the interpretation of 'tired' as indicating TIRED*, is calibrated so as to justify the implicature in (13). The explicature could therefore only be inferred once the implicature had been tentatively assumed to be part of Jane's meaning. The overall interpretation results from a process of mutual adjustment between explicature and implicature.

Consider now an expression such as 'some of the Xs', which is generally seen as giving rise to 'scalar implicatures'. From a semantic point of view, 'some of the Xs' denotes the set of subsets of *n* Xs where *n* is at least two and at most the total number of Xs. From a relevance-theoretic point of view, an expression of the form 'some of the Xs' – like any linguistic expression – is used not to encode the speaker's meaning, but to indicate it. In particular, the concept indicated by a given use of 'some of the Xs' may be an *ad hoc* concept SOME OF THE XS* whose denotation differs from that of the literal SOME OF THE XS. Rather than ranging over all subsets of Xs between two and the total number of Xs, the extension of SOME OF THE XS* may be narrowed at either end, or it may be broadened to include subsets of one.

Imagine (14) uttered in a discussion of the spread of scientific knowledge in America:

(14) Most Americans are creationists and some even believe that the Earth is flat.

Clearly, the speaker is understood as meaning that a number of Americans much greater than two believe that the earth is flat. Two Americans with this belief – say two inmates in a psychiatric hospital – would be enough to make her utterance literally true, but not (and by a wide margin) to make it relevant. Since we can assume that the speaker regards it as common knowledge that not all Americans believe the earth is flat, there is no reason to think that this is part of her meaning (inferring it would involve a processing cost without increasing cognitive effects, so it would detract from relevance). On the other hand, the speaker's contrastive use of 'most' and 'some' and her use of 'even' do make it part of her meaning that fewer Americans believe the earth is flat than believe in creationism (this, of course, entails that not all Americans believe that the earth is flat, but not every entailment of a speaker's meaning is part of that meaning). So the denotation indicated by the use of 'some' in (14) is narrower at both ends than the literal denotation: it includes those subsets of Americans which are large enough to be relevant (and hence much larger than sets of two Americans), but smaller than the set of American creationists.

Let us now go back to a version of example (4). Jane and Henry have invited a few friends to a dinner party. Suppose, first, that they have agreed that Henry

will go and pick up the dessert from the *patisserie* as soon as the guests begin to arrive. Henry is in the garage; he hears the bell ring, and then Jane shouts (15):

(15) JANE [*to Henry*]: Some of the guests have arrived

Henry does not know how many of the guests have arrived, or indeed whether Jane has opened the door and seen how many there are, and the question need not even occur to him. What makes Jane's utterance relevant is that it implies that he should go and buy the dessert now, and this does not depend on the number of guests at the door. Henry's construal of 'some' is compatible with any number of guests having arrived, even a single one, and it therefore involves a broadening of the literal meaning.

Consider now a different scenario. Henry is alone in the kitchen cooking. Jane comes in and tells him (15). The implications that Henry derives are that he should come and greet the guests and bring the finger food he has made as an appetizer. The value of 'some' is taken to be one for which these are the main consequences. If all the guests had arrived, the implications would be not just that he should greet the guests and bring the finger food, but also, and more importantly, that he should put the fish in the oven and make the final preparations for the meal itself. The fact that Jane's utterance achieves relevance without bringing to mind consequences more typical of the arrival of all the guests causes Henry to construe 'some' with some vague cardinality above one and below all. He need not actively exclude *all*; he may simply not even consider it. On the other hand, if he had been wondering whether all the guests have arrived, then he will take Jane's utterance to license the inference that not all of them have. Moreover, if he had asked Jane whether all the guests had arrived, or if he knew she was aware that it was particularly relevant to him at this point in time, he would take that to be an intended inference. The same would happen if she had put a contrastive stress on 'some', causing him extra effort and suggesting an extra effect. In other words, if there is some mutually manifest, actively represented reason to wonder whether all the guests have arrived, then (15) can be taken to implicate that not all of them have.

From a relevance theory point of view, (11), (14) and (15) are just ordinary illustrations of the fact that linguistic expressions serve to indicate rather than encode the speaker's meaning, and that the speaker's meaning is quite often a narrowing or broadening of the linguistic meaning. Taking 'some' to indicate not *at least two and possibly all* but *at least two and fewer than all* is a common narrowing of the literal meaning of 'some' at the level of the explicature of the utterance. It is not automatic, but takes place when the implications that make the utterance relevant as expected are characteristically carried by this narrowed meaning.

We are not denying that a statement of the form '. . . some . . .' may in some cases carry an implicature of the form . . . *not all* . . . (or, in other cases we will not discuss here, an implicature of the form . . . *some* . . . *not* . . .). This happens

when the utterance containing 'some' achieves relevance by answering a tacit or explicit question about whether *all* items satisfy the predicate. The fact that it does not give a positive answer *implicates* a negative answer, and therefore a narrowed construal of 'some' as excluding all. Standard accounts of 'scalar implicatures' fail to distinguish between cases where the explicature merely entails . . . *not all* . . . and the much less frequent cases where the utterance also implicates . . . *not all* . . ..

In all cases where the meaning of 'some' in an utterance is narrowed to exclude *all*, this is the result of an inferential process which looks at consequences that might make the utterance relevant as expected, and which adjusts the meaning indicated by 'some' so as to yield these consequences. In particular, if what would make the utterance relevant is an implication that is true of some but not all Xs, then the meaning of 'some' is adjusted to exclude *all*. These inferential processes result from the hearer's automatic search for an interpretation that meets his expectation of relevance, and they all follow the same heuristics. There is nothing distinctive about the way 'scalar' inferences are drawn. Moreover, the class of cases described in the literature as scalar inferences is characterised by an enrichment at the level of the explicature (where, for instance, 'some' is reinterpreted in a way that excludes *all*), and only in a small sub-class of these is the exclusion of the more informative concept not just entailed but also implicated.

According to relevance theory, then, so-called 'scalar implicatures' are neither scalar nor necessarily implicatures. Of course, it would be possible to redefine the notion of 'scalar implicature' to cover just those cases where there is an explicit or implicit question about whether the use of a more informative expression by the speaker (e.g. 'all' instead of 'some') would have been warranted; here, a denial of the more informative claim can indeed be implicated by use of the less informative expression. However, 'scalar implicatures' in this restricted sense depend on contextual premises (linked to the fact that the stronger claim was being entertained as a relevant possibility) rather than a context-independent scale, and are therefore not candidates for the status of GCI.

From the point of view of relevance theory, then, the classical neo-Gricean theory of scalar implicatures can be seen as a mistaken generalisation of the relatively rare case where a weaker claim genuinely *implicates* the denial of a stronger claim which is under consideration in the context, to the much more common case where the denotation of an expression is narrowed to exclude marginal or limiting instances with untypical implications. For instance, 'possible' as in (3a) ('It is possible that Hillary will win') is often construed as excluding, on the one side, mere metaphysical possibility with a very low empirical probability, and, on the other, certainty and quasi-certainty. The trimming of 'possible' at both ends results in an enriched and generally more

relevant meaning. Since the trimming at the very high probability end is no different from the trimming that takes place at the very low probability end, both should be explained in the same way. This rules out the scalar aspect of the 'scalar implicature' account, which works (if at all) only at the upper end. By contrast, if (3a) were uttered in reply to the question: 'Is it certain that Hillary will win?', then it would indeed implicate (3c) ('It is not certain that Hillary will win'), because it would achieve relevance by implicitly answering in the negative a question that had been asked. From a relevance theory point of view, the two cases should be distinguished.

This is not the place to compare in detail the GCI and relevance-theoretic approaches. Instead, we focus on a testable difference in their predictions. According to Levinson, 'GCI theory clearly ought to make predictions about process. But here the predictions have not yet been worked out in any detail' (Levinson 2000: 370). However, there is one prediction about process that follows quite directly from GCI theory, since it amounts to little more than a restatement of some of the tenets of the theory. According to the theory, GCIs are computed by default, and are contextually cancelled when necessary. Both the computation and the cancellation of GCIs are processes, and each should therefore take some time and effort (even if the default nature of GCIs should make their computation quite easy and rapid). Everything else being equal, less effort should be required, and less time taken, in the normal case where a GCI is computed and not cancelled, than in the exceptional case where a GCI is first computed and then cancelled. Relevance theory predicts just the opposite pattern.

From a relevance-theoretic perspective, the speaker's meaning is always inferred, even when it involves a literal interpretation of the linguistic expressions used. However, the inferences may differ in the time and effort they require. Both sentence meaning and context contribute to making some interpretations easier to derive than others. If sentence meaning were the only factor to be taken into account, one could predict that the smaller the distance between it and the speaker's meaning it is used to indicate, the less time and effort would be required to bridge the gap between sentence meaning and speaker's meaning. However, contextual factors must also be taken into account. For instance, an enriched interpretation may be primed by the context, and may therefore be easier to infer than a literal interpretation. Consider a variant of example (11):

(16) HENRY: You look tired. Let's go to the cinema.
JANE: I am tired, but not too tired to go on working.

A natural interpretation of Henry's utterance involves the *ad hoc* concept TIRED*, where being TIRED* is a sufficient reason to stop working but not a sufficient reason to stay at home. Jane could have replied, 'No, I am not tired: I'll go on working', meaning that she was not TIRED* (as discussed above).

Table 14.1. *Contrasting predictions of GCI Theory and relevance theory about the speed of interpretation of scalar terms (when an enriched construal is not contextually primed)*

| Interpretation of the scalar term | GCI theory | relevance theory |
|---|---|---|
| literal | default enrichment + context-sensitive cancellation, ***hence slower*** | no enrichment, ***hence faster*** |
| enriched | default enrichment, ***hence faster*** | context-sensitive enrichment, ***hence slower*** |

When Jane asserts instead that she *is* tired, Henry is primed to interpret 'tired' as TIRED*. However, a relevant interpretation of Jane's utterance as a whole imposes a broader, more literal and, in this situation, more effortful construal of the term.

Even when an enriched interpretation of an utterance is not primed by the context, it may require less processing effort than the literal interpretation, because the contextual implications that make the enriched interpretation relevant are easier to derive than those that would make the literal interpretation relevant. This typically occurs with metaphorical utterances, where a relevant literal interpretation is often hard, or even impossible, to construct.

In the absence of contextual factors that would make an enriched interpretation of an utterance easier to arrive at, relevance theory predicts that a literal interpretation, which merely involves the attribution to the speaker of a meaning already provided by linguistic decoding, should require shallower processing and take less time than an enriched one, which involves a process of meaning construction. This is the case in particular in the experiments we describe below.

The difference in predictions between GCI theory and relevance theory can be presented in table form (see Table 14.1). This difference is of a type that lends itself to experimental investigation.

## 14.4 Methodological considerations in experimental approaches to 'scalar inferences'

In the experimental study of scalar inferences,[1] there are four methodological considerations to bear in mind. First, one wants to be sure that a given result (e.g. the rate of responses indicating a pragmatic enrichment, or the mean reaction time associated with such an enrichment) is a consequence of the intended target of the experiment and not of other contextual variables. For

example, one wants to be sure that the understanding of a disjunctive statement of the form *P or Q* as excluding *P and Q* is due to pragmatic enrichment of the term 'or' (from an inclusive to an exclusive interpretation) rather than to some other feature. It is therefore best to avoid investigating utterances which invite an exclusive understanding of the situation described, as opposed to an exclusive understanding of the description itself. In example (6) above ('Jane is in Paris or in Madrid'), the exclusive understanding is based on our knowledge that a person cannot be in two places at once, and need not involve any pragmatic enrichment of the meaning of the word 'some'. In devising experimental material, it is thus important to invent examples where an enriched interpretation is not imposed by extra-pragmatic considerations. This can be done by using examples where the participants' knowledge is equally compatible with a literal or an enriched interpretation of a scalar term, or where knowledge considerations might bias participants in favour of a literal interpretation. In either case, if the results provide evidence of enrichment, one can be confident that it comes from a pragmatic inference about what the utterance meant, rather than a mere understanding of how the world is.

Second, it is best to use a paradigm that allows for two identifiable outcomes, so that the presence of an enrichment can be indicated by a unique sort of response, while a non-enrichment is indicated by a different response. This is why most of the experiments on scalars described here involve a scenario that could be described by use of a more informative utterance than the test utterance (produced by a puppet or some other interlocutor). Imagine, for example, being shown five boxes, each containing a token, and being told, 'Some boxes contain a token'. If you interpret 'some' literally (i.e. as compatible with *all*), you would agree with the statement; if you enrich 'some' so as to be incompatible with *all*, you would have to disagree. In these conditions, a participant's response (agrees or disagrees) is revealing of a particular interpretation.

Third, one wants every assurance that an effect is robust. That is, one wants to see the same result over and over again, across a variety of comparable tasks. When two similar studies (for instance, two studies investigating different scalar terms, but in equivalent ways) produce comparable outcomes, each strengthens the findings of the other. By contrast, if two very similar experiments fail to produce the same general effects, something is wrong. This does not mean that negative results are necessarily fatal for an experimental paradigm. A carefully modified experiment which prompts a different sort of outcome than previous ones (and in a predictable way) can help determine the factors underlying a certain effect. This happens with the developmental findings to be described below, which have generally shown that children are more likely than adults to *agree* with a weak statement (e.g. 'Some horses jumped over a fence') when a stronger one would be pragmatically justified (because in fact all the horses jumped over a fence). All sorts of follow-up studies have been designed to put

this effect to the test. In general, the effect has been resilient, but there are a few studies showing that one can get children to appear more adult-like by using specific sorts of modifications. For example, experimenters have tried to confirm the effect in conditions where participants are given some prior training, or using scenarios designed to highlight the contrast between the weak utterance and the possibility of making a stronger claim. The net result is that the outcomes of these tests do indeed help identify the factors that can encourage scalar inference-making.

Fourth, it is important for any experiment to include as many reasonable controls as possible. These are test questions which are similar to the main items of interest, but are used basically to confirm that there is nothing bizarre in the task. For example, if participants' responses indicate that they enrich 'some', but it is also found that the same participants endorse the use of the word 'some' to describe a scene where 'none' would be appropriate, then there is something questionable about the experiment. This rarely happens (the above example is presented for illustrative purposes only), but it is important to provide assurances for oneself and for readers that such bizarreness can be ruled out. Any decent task will include several controls which lead to uncontroversial responses and are designed, in effect, to contextualise the critical findings. The studies we will discuss exemplify the four methodological considerations we have just discussed.

## 14.5 Developmental studies

The experimental study of scalar inferences began in the framework of developmental studies on reasoning. Noveck (2001) investigated the responses children gave (by agreeing or disagreeing) to a puppet who produced several statements, including one that could ultimately lead to a pragmatic enrichment. All the statements, even those used as controls to confirm that the participants understood the task, concerned the contents of a covered box, and were presented by a puppet (handled by the experimenter). Participants were told that the contents of the covered box resembled those of one or other of two further boxes, both of which were open and had their contents in full view. One open box contained a parrot, and the other contained a parrot and a bear. The participants then heard the puppet say:[2]

(17) A friend of mine gave me this (covered) box and said, 'All I know is that whatever is inside this box (the covered one) looks like what is inside this box (the one with a parrot and bear) or what is inside this box (the one with just a parrot)'.

The participant's task was to say whether or not he agreed with further statements produced by the puppet. The key item was ultimately the puppet's 'underinformative' statement:

(18) There might be a parrot in the box.

Given that the covered box *necessarily* contained a parrot, the statement in (18) can be answered in one of two ways. The participant can 'agree' if she interprets 'might' literally (so that . . . *might* . . . is compatible with . . . *must* . . .) or she can 'disagree' if she interprets *might* in an enriched way (where . . . *might* . . . is incompatible with . . . *must* . . .). Adults tended to be equivocal with respect to these two interpretations (35% agreed with the statement), while children (five-, seven- and nine-year-olds) tended to interpret this statement in a minimal way, i.e. literally. Collectively, 74 per cent of the children responded by agreeing with the statement in (18). However, not all children were alike.

The five-year-olds agreed with (18) at a rate of 72 per cent (a percentage unlikely to occur by chance – which would yield a rate of 50 per cent in such agree/disagree contexts). Nevertheless, they failed to answer many control questions at such convincing rates. For example, when asked to agree or disagree with statements about the bear ('There has to be a bear', 'There might be a bear', 'There does not have to be a bear', 'There cannot be a bear') they answered at levels comparable to those predicted by chance (55% correct across the four questions). Seven-year-olds, on the other hand, did manage to answer practically all seven control questions at rates indicating that they understood the task overall (77%). This is why Noveck (2001) reported that seven-year-olds were the youngest to demonstrate competence with this task while at the same time revealing that they preferred the literal interpretation of 'might' (at a rate of 80%, which is statistically distinguishable from expectations based on chance). The seven-year-olds thus provided the strongest evidence that those linguistically competent children who performed well on the task overall still interpreted 'might' in an unenriched way. As might be expected, the nine-year-olds also answered control problems satisfactorily. Response rates indicating unenriched interpretations of 'might' were high (69%), and much higher than the adults', but were nevertheless statistically indistinguishable from predictions based on chance, which suggests that these children were *beginning* to appear adult-like with respect to (18). Overall, these results were rather surprising for a reasoning study, because they indicated that children were more likely than adults to produce a logically correct evaluation of the under-informative modal statement. This sort of response is surprising and rare, but thanks to a pragmatic analysis – where pragmatically enriched interpretations are seen as likely to result from a richer inferential process than minimal interpretations that add nothing to semantic decoding – these results had a ready interpretation.

Despite taking every precaution (using numerous control items and sampling many children), one can never exclude the possibility that these effects might be a result of some subtle factor beyond the experimenter's intention or control.

That is why – especially when faced with counterintuitive results like these – it pays to do follow-ups. These have essentially been of two sorts.

The first sort of follow-up is designed to confirm that the effect exists. In one experiment (Noveck 2001: Experiment 2), five-year-olds, seven-year-olds and adults were given the same task as the one above, but all participants received more thorough training to ensure that they understood the parameters of the task. The training involved an identical scenario (one box containing a horse and a fish and another just a horse), but participants were asked pointed questions about the covered box (e.g. *Could there be a fish by itself in the box?*). Overall, such training increased rates of minimal interpretations of 'might' across all three ages when participants were given the task in Experiment 1. Agreement with a statement such as (18) was now 81 per cent for five-year-olds, 94 per cent for seven-year-olds, and 75 per cent for adults. Although rates of such minimal interpretations were statistically comparable across ages, the same trends are found as in the first experiment reported above. Seven-year-olds again demonstrated (through their performance with the control problems) that they were the youngest to show overall competence with the task while *tending* to be more likely than adults to choose a literal interpretation of the weak scalar term. The data also revealed that the extra training encourages adults to behave more 'logically' (to stick to the literal meaning of 'might'), like the children.

In an attempt to establish the reliability and robustness of the developmental effect, Noveck (2001: Experiment 3) took advantage of an older study which unintentionally investigated weak scalar expressions in four- to seven-year-old children and which also failed to show evidence of pragmatic enrichment. Smith (1980) presented children with statements such as 'Some giraffes have long necks' and reported that it was surprising to find the children accepting them as true. In a third experiment, therefore, Noveck (2001) essentially continued from where Smith left off. The experiment adopted the same technique as Smith (which included pragmatically felicitous statements such as 'Some birds live in cages' as well as statements with 'all') in order to confirm that the developmental findings of the first two experiments were not flukes. The only differences in this third experiment were that the children were slightly older than in the first two studies (eight and ten years old), and that the experimenter was as 'blind' to the purpose of the study as the participants (the student who acted as experimenter thought that unusual control items such as 'Some crows have radios' or 'All birds have telephones' were the items of interest). The results showed that roughly 87 per cent of children accepted statements like 'Some giraffes have long necks', whereas only 41 per cent of adults did. Again, adults were more likely than children to enrich the interpretation of the under-informative statements (understanding . . . *some* . . . to exclude . . . *all* . . .) and thus tended to reject them (since all giraffes have long necks). All participants answered the five sorts of control items (25 items altogether) as one would expect.

These data prompted Noveck (2001) to revisit other classic studies that serendipitously contained similar scenarios (where a stronger statement would be appropriate but a weaker one is made) to determine whether they tell the same story as 'might' and 'some'. In fact, three studies with 'or' (Paris 1973; Sternberg 1979; Braine and Rumain 1981), where a conjunctive situation is described with a weaker disjunction, provide further confirming evidence. The authors of these studies also reported counter-intuitive findings which show younger children being, in effect, more logical than adults (children tend to treat 'or' inclusively more often than adults). None of these authors, lacking a proper pragmatic perspective, were able to make sense of these data at the time. All told, this effect appeared robust.

Other follow-up studies have actually taken issue with Noveck's *interpretation* of the findings. In fact, Noveck (2001: 184) emphasised that his data show that children are ultimately less likely than adults to pragmatically enrich underinformative items across tasks; this did not amount to a claim that children lacked pragmatic competence. Still, there has been a lot of work designed to show that young children are more competent than it might appear. These studies usually take issue with Noveck's Experiment 3 (the one borrowed from Smith 1980), because it involves the quantifier 'some' (which is of more general interest than 'might'), and because the items used in that task are admittedly unusual (see Papafragou and Musolino 2003; Chierchia, Guasti, Gualmini, Meroni, Crain and Foppolo 2004; Feeney, Scrafton, Duckworth and Handley 2004; Guasti, Chierchia, Crain, Foppolo, Gualmini and Meroni 2005).

We highlight here the main advances made in these studies. In two sets of studies, Papafragou and colleagues (Papafragou and Musolino 2003; Papafragou and Tantalou 2004) attempted to show that children as young as five are generally able to produce implicatures if the circumstances are right. In fact, Papafragou and Musolino (2003: Experiment 1) first confirmed the developmental effect summarised above by showing that five-year-olds are less likely than adults to produce enrichments with 'some', 'start' and 'three' in cases where a stronger term (namely, 'all', 'finished' and a 'larger number', respectively) was called for. They then modified the experimental setup in two ways in preparing their second experiment. First, before they were tested, participants received training designed to enhance their awareness of pragmatic anomalies. Specifically, children were told that the puppet would say 'silly things' and that the point of the game was to help the puppet say it better (e.g. they would be asked whether a puppet described a dog appropriately by saying 'This is a little animal with four legs'). In the event that the child did not correct the puppet, the experimenter did. Second, the paradigm put the focal point on a protagonist's performance. Unlike in their Experiment 1, where participants were asked to evaluate a quantified statement like 'Some horses jumped over the fence' (when in fact all the horses did), the paradigm in Experiment 2 creates the

expectation that the stronger statement (with 'all') might be true. Participants would hear a test statement like, 'Mickey put some of the hoops around the pole' (when he had been shown to succeed with all of the hoops), and they were also told that Mickey claims to be especially good at this game and that this is why another character challenges him to get all three hoops around the pole. With these changes, five-year-olds were more likely to produce enrichments than they were in the first experiment. Nevertheless, the five-year-olds, even in the second experiment, still produced enrichments less often than adults did. This indicates that – even with training and with a focus on a stronger contrast – pragmatic enrichments require effortful processing in children.[3]

Guasti, Chierchia, Crain, Foppolo, Gualmini and Meroni (2005) argue that pragmatic enrichments should be as common among five-year-olds as among adults, and further investigated the findings of Noveck (2001) and Papafragou and Musolino (2003). In their first experiment, they replicated the finding of Noveck (2001: Experiment 3) on 'some' in seven-year-olds, and used this as a baseline for studying independently the role of the two factors manipulated by Papafragou and Musolino (2003). One factor was the role of training and its effect on children's proficiency at computing implicatures (Experiments 2 and 3), and the other was the role of increasing emphasis on the outcome of a scalar implicature (Experiment 4). Their Experiments 1 through 3 showed that training young participants to produce the most specific description of a given situation can indeed have a major effect on performance. While their initial experiment showed that seven-year-olds accept statements such as 'Some giraffes have long necks' 88 per cent of the time (as opposed to 50% for adults), when trained in this way their acceptance rate becomes adult-like and drops to 52 per cent. Nonetheless, this effect is short-lived, i.e. it does not persist when the same participants are tested a week later (Experiment 3). In the last experiment, the authors made the *all* alternative more salient in context. They did this, for instance, by presenting participants with a story in which several characters have to decide whether the best way to go and collect a treasure is to drive a motorbike or ride a horse. After some discussion, all of them choose to ride a horse. In this way, it is made clearer that the statement participants have to evaluate ('Some of the characters chose to ride horse') is under-informative. The results indicated that children are more likely to produce an enriched interpretation in an adult-like manner when the context makes this enrichment highly relevant.

This last finding shows that one can create situations that encourage children to pragmatically enrich weak-sounding statements, and to do so in an adult-like way. It does not alter the fact that in less elaborate scenarios, where cues to enrichment are less abundant, seven-year-olds do not behave in this way, and it does not tell us what younger children do. Overall, the developmental effect shows that pragmatic enrichments require some effort. In experimental settings,

the effort required can be somewhat reduced, or the motivation to perform it increased, but in the absence of such contextual encouragements, younger children faced with a weak scalar term are more likely to stick with its linguistically encoded meaning.

If children had been found to perform scalar inferences by default, this would have been strong evidence in favour of the GCI theory approach. However, taken together, the developmental data suggest that for children, enriched interpretations of scalar terms are not default interpretations. This sort of data is not knock-down evidence against GCI theory, since it is compatible with two hypotheses: (1) scalar inferences are not default interpretations for adults either (even if adults are more likely to derive them because they take relatively less effort, and because adults are more inclined to invest effort in the interpretation of an utterance given their greater ability to derive cognitive effects from it). Or, (2) in the course of development, children become not only capable of performing scalar inferences by default, but also disposed to perform such inferences. The first hypothesis is consistent with the relevance theory approach, while the second is consistent with the GCI approach. To find out which approach has more support, further work had to be done with adults.

## 14.6 Time course of comprehension among adults

As mentioned above, GCI theory implies that a literal interpretation of a scalar term, produced by cancelling a default enrichment, should take longer than an enriched interpretation; by contrast, relevance theory, which denies that enrichment takes place by default, implies that an enriched interpretation, inferred when required to meet contextual expectations of relevance, should take longer than a literal one. What is needed to test these contrasting predictions are experiments manipulating and measuring the time course of the interpretation of statements with weak scalar terms.

The same methodological considerations apply here as in the developmental tasks: it is important to make sure that enriched interpretations are clearly identifiable through specific responses, that the tasks used include a variety of controls, and that the effect is reliable and robust. One way of identifying enriched vs. literal interpretations is provided by earlier studies where participants were asked to make true/false judgements about statements (e.g. 'Some elephants are mammals') which could be construed as literally true but underinformative, or enriched (to imply . . . *not all* . . .) and judged as false. Hence the participants' truth-value judgements reflect their literal or enriched interpretation.

As indicated above, prior work is often critical to developing the appropriate measures. In fact, Rips (1975) unintentionally included the right sort of cases when looking at other issues of categorisation using materials such as 'Some

congressmen are politicians'. He examined the effect of the interpretation of the quantifier by running two studies, one where participants were asked to treat 'some' as meaning *some and possibly all*, and another where they were asked to treat 'some' as meaning *some but not all*. The results showed that participants given the *some but not all* instruction in one experiment responded more slowly than those given the *some and possibly all* instruction in another. Despite these indications, Rips modestly hedged in concluding that 'of the two meanings of *Some*, the informal meaning *may* be the more difficult to compute' (italics added). To make sure that Rips's data were indeed indicative of a slowdown related to *some but not all* readings, Bott and Noveck ( ) ran a series of four experiments that followed up on Rips (1975) and essentially confirmed that enriched interpretations take longer than literal ones.

Bott and Noveck's categorisation task involved the use of under-informative items (e.g. 'Some cows are mammals') and five controls that varied the quantifier (*some* and *all*) and the category–subcategory order, as well as proper membership. The six types of statements are illustrated below with the six possible ways of using the subcategory *elephants*, but it is worth pointing out that the paradigm was set up so that the computer randomly paired a given subcategory with a given category while verifying that, at the end of each experimental session, there were nine instances of each type:

(19) a. Some elephants are mammals (Under-informative).
b. Some mammals are elephants.
c. Some elephants are insects.
d. All elephants are mammals.
e. All mammals are elephants.
f. All elephants are insects.

In the first experiment, a sample of twenty-two participants was given the same task twice, once with the instruction to treat 'some' as meaning *some and possibly all*, and once with the instruction to treat 'some' as meaning *some but not all* (and of course the order of presentation was varied). When participants were under instruction, in effect, to engage the scalar inference, they were shown to be less accurate and take significantly longer to respond to the Under-informative items (like those in (19a)). Specifically, when the instructions called for a *some but not all* interpretation, rates of correct responses to the Under-informative item (i.e. judging the statement 'false') were roughly 60 per cent; when the instructions called for a *some and possibly all* interpretation, rates of correct responses to the Under-informative item (i.e. judging the statement 'true') were roughly 90 per cent. For the control items, rates of correct responses were always above 80 per cent and sometimes above 90 per cent. It is clear that the Under-informative case in the *some but not all* condition provides exceptional data.

The reaction time data showed that the correct responses to the Under-informative item in the *some but not all* condition were exceptionally slow. It took roughly 1.4 seconds to correctly evaluate the Under-informative statements in the *some but not all* condition and around 0.8 seconds in the *some and possibly all* conditions. Responses to the control items – across both sorts of instructions – took at most 1.1 seconds, but more often around 0.8 to 0.9 seconds. Thus, the Under-informative statement in the *some but not all* condition is the one most affected by the instructions. All this confirms Rips's initial findings. More importantly, there is not a single indication that interpreting 'some' to mean *some but not all* is an effortless or quasi-effortless step. Again, a default view of scalar inference would predict that under the *some but not all* instruction, responses to Under-informative statements should take less time than responses under the *some and possibly all* instruction. According to an account based on relevance theory, the opposite should be found. The data more readily support the relevance-theoretic account.

A potential criticism of this experiment might go as follows. Given that the correct response to the Under-informative statement with the *some and possibly all* instruction is to say 'True', while the correct response to the Under-informative statement with the *some but not all* instruction is to say 'False', the reduced accuracy and slowdown in reaction times in the second type of case might be due to a response bias favouring positive rather than negative responses. To alleviate concerns about such a potential response bias, Bott and Noveck demonstrated experimentally that the effects linked to pragmatic effort are not simply due to hitting the 'False' key.

In a second experiment, the paradigm was modified so that the same overt response could be compared across both sorts of instructions; that way, the participants' response choice (True vs. False) could not explain the observed effects. To make these comparisons possible, participants were not asked to agree or disagree with first-order statements such as those in (19), but with second-order statements about these first-order statements. For example, participants were presented with the two statements: 'Mary says the following sentence is false' / 'Some elephants are mammals.' They were then asked to agree or disagree with Mary's second-order statement. In this case, participants instructed to treat 'some' as meaning *some but not all* should agree, whereas participants instructed to treat 'some' as meaning *some and possibly all* should disagree, reversing the pattern of positive and negative responses in the previous experiment.

The results of this second experiment were nevertheless remarkably similar to those of the first one. Here, when participants were, in effect, under instruction to draw the scalar inference, they were less accurate and took significantly longer to respond correctly to the Under-informative item. When 'agree' was linked with the instruction to use a *some but not all* interpretation, rates of

correct responses were roughly 70 per cent; when 'agree' was linked with the instruction to use a *some and possibly all* interpretation, rates of correct responses were roughly 90 per cent. For all control items, rates of correct responses were always above 85 per cent, and often above 90 per cent. It is clear that, once again, the Under-informative case in the *some but not all* condition provides exceptional data. The reaction-time data also showed that the correct 'agree' responses to the Under-informative item in the *some but not all* condition were exceptionally slow. It took nearly 6 seconds to evaluate the Under-informative statements correctly when 'agree' was linked with the instruction to use a *some but not all* interpretation, and around 4 seconds when 'agree' was linked with the instruction to use a *some and possibly all* interpretation (all reaction times were longer than in the previous experiment due to the *Mary says* statement). The control items across both sorts of instructions took on average around 4.5 seconds, and never more than 5 seconds. Again, the experiment demonstrated that any response that requires a pragmatic enrichment implies extra effort.

Both these experiments, though inspired by previous work, are arguably unnatural. It is unusual to instruct participants in a conversation about how they should interpret the word 'some', as was done in Experiment 1; the second experiment doubles the complexity by requiring participants to make metalinguistic judgements based on statements like *Mary says the following is false*. Bott and Noveck's third experiment simplified matters by asking participants to make true/false judgements about the categorical statements (e.g. those in (19)) themselves, and with no prior instruction. When the issue is presented in this way, there is no useful sense in which a response is 'correct' or not. Rather, the responses reveal the participant's literal or enriched interpretation, and can be compared in terms of reaction times.

Roughly 40 per cent of participants responded 'true' to Under-informative items and 60 per cent responded 'false'. This corresponds to the rates found among adults in Noveck's developmental studies (see also Noveck and Posada 2003; Guasti, Chierchia, Crain, Foppolo, Gualmini and Meroni 2005). The main finding was that mean reaction times were longer when participants responded 'false' to the Under-informative statements than when they responded 'true' (3.3 seconds versus 2.7 seconds, respectively). Furthermore, 'false' responses to the Under-informative statements appear to be slower than responses to all the control statements (including three, (19c), (19e) and (19f), that require a 'false' response). The 'true' response was made at a comparable speed to all of the control items.

In their last experiment, Bott and Noveck varied the time available to participants for responding to the statements. The rationale for this design was as follows: if, as implied by GCI theory, literal interpretations of weak scalar terms take longer than the default enriched interpretations, then limiting the

time available should decrease the rate of literal interpretations and increase the rate of enriched ones. By contrast, if, as implied by relevance theory, enriched interpretations take longer, then limiting the time should have the opposite effect (i.e. shorter lags should be associated with higher rates of literal interpretations). Using the same general procedure as in the previous experiments (asking participants to make true/false judgements about categorical statements), the paradigm manipulated the time available for the response. In one condition, participants had a relatively short time (0.9 seconds) to respond, while in the other they had a relatively longer time (3 seconds). Only the time to *respond* was manipulated. To control for uptake, participants were presented with the text one word at a time, and at the same rate in both conditions; there is thus no possibility that participants in the 'Short-lag' condition spent less time reading the statements than those in the 'Long-lag' condition.

Bott and Noveck reported that when participants had a shorter period of time available to respond, they were more likely to respond 'True' to Under-informative statements (indicating a literal interpretation): 72 per cent of participants responded 'True' in the Short-lag condition, and 56 per cent did so in the Long-lag condition. This strongly implies that they were less likely to derive the scalar inference when they were under time pressure than when they were relatively pressure-free. As in all the prior experiments, control statements provide a context in which to appreciate the differences found among Under-informative statements. The results showed that performance on control statements in the Short-lag condition was quite good overall (rates of correct responses ranged from 75% to 88%) and that, as one would expect, rates of correct performance on the control items *increased* when more time was available (by 5% on average). The contrast between a percentage that drops with extra time (as is the case for the Under-informative statements) and percentages that increase over time provides a unique sort of interaction, confirming that time is needed to provoke scalar inferences.

The experiments we have described so far take into account the four methodological considerations discussed above, with well-controlled dependent variables: the rate of literal vs. enriched interpretations of weak scalar terms, and the speed with which they are derived. Together, they provide strong evidence that an enriched interpretation of a weak scalar term requires more processing time than an unenriched, literal interpretation, as predicted by relevance theory and contrary to the prediction implied by GCI theory.

Still, it might be argued that the categorisation tasks used in these experiments, even if they are methodologically sound from an experimental psychology point of view, are too artificial to be used in testing pragmatic hypotheses. If the claim were that laboratory tasks are somehow irrelevant to pragmatics, we would argue that the onus of the proof is on the critics: after all, participants bring their ordinary pragmatic abilities to bear on experimental verbal tasks, just

as they do in any unusual form of verbal exchange. In particular, if it is part of adult pragmatic competence to make scalar inferences by default, it would take some argument to make it plausible that an experimental setting somehow inhibits this basic disposition. On the other hand, if the claim is that fairly artificial laboratory experiments are not enough, and that they should be complemented with more ecologically valid designs, we agree. Happily, Breheny, Katsos and Williams (2006) have provided just this kind of welcome complement.

Following up on a procedure from Bezuidenhout and Cutting (2002), Breheny *et al*. presented disjunctive phrases (e.g. 'the class notes or the summary') in two kinds of contexts: Lower-bound contexts (where the literal reading of a scalar term is more appropriate, as in (20) below), and Upper-bound contexts (where the enriched reading of the scalar is more appropriate, as in (21) below). These were presented as part of short vignettes (along with many 'filler' items to conceal the purpose of the study) and participants' reading times were measured. More specifically, participants were asked to read short texts presented on a computer screen one fragment at a time, and to read the next fragment by hitting the space bar (the slashes in (20) and (21) delimit fragments).

(20) *Lower-bound context*
John heard that / the textbook for Geophysics / was very advanced. / Nobody understood it properly. / He heard that / if he wanted to pass the course / he should read / *the class notes or the summary.*

(21) *Upper-bound context*
John was taking a university course / and working at the same time. / For the exams / he had to study / from short and comprehensive sources. / Depending on the course, / he decided to read / *the class notes or the summary.*

In such a task, if shorter reading times were found in the Upper-bound contexts, which call for scalar inferences, than in the Lower-bound contexts, where the literal interpretation is more appropriate, this would support the GCI claim that scalar inferences are made by default. Findings in the opposite direction would support the relevance theory account. What Breheny *et al*. found is that phrases like *the class notes or the summary* took significantly longer to process in Upper-bound contexts than in Lower-bound contexts, a result consistent with findings reported above.

## 14.7 Conclusion

The experimental work we have summarised here confirms predictions derived from relevance theory, and falsifies predictions derived from GCI theory. Does this mean that relevance theory is true and GCI theory is false? Of course not.

Nevertheless, these results should present a serious problem for GCI theorists. It is quite possible that they will find a creative solution to this problem. For instance, they might be able to show that, despite the methodological precautions described above, the reported studies failed to eliminate some uncontrolled factor, and that better studies provide evidence that points in the opposite direction. More plausibly, they might revise their theory in order to accommodate these results. One line of revision would be to reconsider the idea that GCIs are default inferences (or to water down the notion of default to the point where it no longer has implications for processing time). After all, not all neo-Griceans agree with Levinson's account of GCIs (see in particular Horn 2004, 2006). Still, it is worth noting that, if scalar inferences are not truly default inferences and invariably involve consideration of what the speaker chose not to say, then we are back to the worry that such inferences are excessively effort-demanding. Generally speaking, experimental findings such as those we have summarised here should encourage neo-Griceans to work out precise and plausible implications of their approach at the level of cognitive processing.

Relevance theorists are not challenged in the same way by the work we have described – after all, their prediction is confirmed – but they should bear in mind that the same prediction could be made from quite different theoretical points of view: it follows from relevance theory, but relevance theory does not follow from it. They might then try to develop aspects of these experiments that could give positive support to more specific aspects of the theory. For instance, according to the theory, hearers look for an interpretation that satisfies their expectations of relevance, and the relevance of an interpretation varies inversely with the effort needed to derive it. It should then be possible to make participants choose a more or a less parsimonious interpretation by increasing or decreasing the cognitive resources available to participants for the interpretation process. Bott and Noveck's fourth experiment can be seen as a first suggestive step in this direction.[4]

As we have just explained, we do not expect readers to form a final judgement on the respective merits of GCI theory and relevance theory on the basis of the experimental evidence presented. What we do hope to have done is to convince you that, alongside other kinds of data, properly devised experimental evidence can be highly pertinent to the discussion of pragmatic issues, and that pragmaticists – and in particular students of pragmatics – might benefit greatly from becoming familiar with relevant experimental work, and contributing to it (perhaps in interdisciplinary ventures).