Are Linguists Better Subjects?

Jennifer Culbertson and Steven Gross

Abstract

Who are the best subjects for judgment tasks intended to test grammatical hypotheses? Michael Devitt ([2006a], [2006b]) argues, on the basis of a hypothesis concerning the psychology of such judgments, that linguists themselves are. We present empirical evidence suggesting that the relevant divide is not between linguists and non-linguists, but between subjects with and without minimally sufficient task-specific knowledge. In particular, we show that subjects with at least some minimal exposure to or knowledge of such tasks tend to perform consistently with one another—greater knowledge of linguistics makes no further difference—while at the same time exhibiting markedly greater in-group consistency than those who have no previous exposure to or knowledge of such tasks and their goals.

1 Introduction

Who are the best subjects for judgment tasks intended to test grammatical hypotheses? Michael Devitt ([2006a], [2006b]) argues, on the basis of a hypothesis concerning the psychology of such judgments, that linguists themselves are. According to Devitt, such judgments—even when fairly immediate and unreflective—are theory-laden, so that those who know more and better theory issue more reliable judgments:

> […] the intuitions [i.e., 'fairly immediate unreflective judgments' – p. 95] that linguistics should mostly rely on are those of the linguists themselves because the linguists are most expert [….] As a result of their incessant observation of language, guided by a good theory, linguists are reliable indicators of syntactic reality [….] ([2006b], p. 111)

Below, we present empirical evidence suggesting that the relevant divide is in fact not between linguists and non-linguists, but rather between subjects with and without minimally sufficient task-specific knowledge. In particular, we show that subjects with at least some minimal exposure to or knowledge of such tasks tend to perform consistently with one another—greater knowledge of linguistics makes no further difference—while at the same time exhibiting markedly greater in-group consistency than those who have no previous exposure to or knowledge of such tasks and their goals.

Our remarks are structured as follows. In §2, we supply some background clarifications. In §3, we examine several previous studies that Devitt wrongly thinks support his claim. In §4, we present our own findings. Finally, in §5, we locate Devitt's claim in the context of his larger discussion in order to distance ourselves from some views he's arguing against and to indicate how our results raise a worry for his psychological model.

2 Background and clarification

Linguists make use of various kinds of judgment in supporting their claims: judgments of acceptability, co-reference, ambiguity, entailment, etc. Our experiment focuses on acceptability judgments (see §3 for an experiment concerning co-reference judgments). Acceptability judgments, according to one rough characterization, are judgments about whether some sentence is 'natural and immediately comprehensible […] in no way bizarre or outlandish.' (Chomsky [1965], p. 10) Or, as we put it in our instructions to subjects (see Appendix), they concern 'whether individual sentences sound good or not *to you* [….] whether you would or could say [them] under appropriate circumstances.'

When the linguist's interest is in grammaticality, such judgments are sometimes called '*grammaticality* judgments.' But this label is also used—indeed, more fittingly—for judgments concerning what sentences are permitted, or generated, by a grammar. (Cf. Schütze [1996], pp. 20-7.) To avoid terminological confusion, we refer to the latter as *beliefs concerning grammaticality*. When well-argued, these are the beliefs, or judgments, of our best current science. They advance explanatory hypotheses based on whatever relevant data and theory linguists might possess, including, but not limited to, the evidence provided by subjects' *acceptability* judgments. Thus, a linguist might explain a pattern of (un)acceptability judgments by advancing the hypothesis that processing in the speaker's mind/brain represents, embodies, or respects a grammar with such-and-such properties, from which it follows that such-and-such sentences are (un)grammatical. Note that acceptability judgments and beliefs concerning grammaticality can diverge. One oft-cited example: subjects, including linguists, judge sentences containing multiple center-embeddings ('The mouse the cat the dog saw chased ate') unacceptable; but linguists judge them grammatical on theoretical grounds, explaining their unacceptability by appeal to memory limitations.[1]

Syntacticians have long relied upon their *own* acceptability judgments as a primary source of evidence. And for just as long this reliance has been questioned. Acceptability judgments—no matter who makes them—require meta-linguistic sophistication and recruit a wide variety of cognitive capacities (including perhaps "folk" beliefs about language and internalized prescriptions) whose relation to the language faculty remains obscure. Moreover, the acceptability judgments linguists often use—their own—tend to be minimally controlled, drawn from a sample of one, and arguably subject to theory-driven bias and/or satiation.

Experimental design, however, always involves complex trade-offs. Thus one might reply to the worries just rehearsed, first, that it is often difficult to find less indirect methods that bear on the relevant syntactic topic; and, second, that controlled studies involving many disinterested subjects are time-consuming and costly to design, implement, and analyze. In an ideal, resource-rich world, in which a greater array of experimental probes had been devised, perhaps no linguist would rely solely on her own acceptability judgments. But as things stand this can seem a reasonable approach, at least in some instances.[2]

Devitt’s argument introduces a further crucial consideration: whether some subjects—linguists, in particular—might be more *reliable*. This consideration, even if correct, would still need to be weighed against others. For instance, even if individual linguists were more reliable than individual non-linguists, the judgments of large numbers of non-linguists might be more reliable than any individual (including individual linguists). And recruiting large numbers of non-linguist subjects is easier than recruiting large numbers of linguists. Still, even if the greater reliability of linguists would not by itself settle whether their use should be preferred, it certainly would provide an important consideration.

But what is it for a subject to be reliable? Devitt clearly uses the term ‘reliable’ in the standard philosophical sense: a process or mechanism is reliable with respect to some target if it tends to yield an accurate indication of that target’s state. Thus, a reliable mercury thermometer is one whose mercury-level tends to be correlated with ambient temperature; a belief-forming mechanism is reliable if the beliefs it produces tend to be true; and, according to Devitt, ‘linguists [on account of their “intuitions”] are reliable [that is, in general accurate] indicators of syntactic reality’ ([2006b], p. 111, cf. p. 104). Psychologists—including psycholinguists—sometimes use the term ‘reliable’ rather for subjects whose responses to a stimulus remain constant across different elicitations, perhaps in different circumstances, regardless of accuracy (e.g. Schütze [1996], p. 187). To avoid confusion, we refer to this property and the related property of agreement across subjects as *consistency*.

Now, in assessing whether linguists’ judgments are indeed more reliable indicators of syntactic reality, it is crucial not to assimilate acceptability judgments and beliefs concerning grammaticality. The claim that linguists are more reliable only has *interest* if it concerns acceptability judgments. That linguists' beliefs concerning *grammaticality* are more reliable would be neither surprising nor of methodological interest. This is obvious if *considered* beliefs concerning grammaticality are at issue. It's not surprising that linguists are more reliable than non-linguists when it comes to doing linguistics—in particular, when it comes to proffering considered explanatory pronouncements in their area of expertise. Nor is this of methodological interest: these proposed explanations are not themselves treated as evidence for further hypotheses (except perhaps in a weak sense supplied by the constraint of coherence); so, the linguists' greater reliability has no upshot for experimental design. But what of *intuitive* (fairly unreflective and immediately arrived at) beliefs concerning grammaticality? Here too it would be no surprise to learn that linguists perform more reliably than non-linguists—indeed, perhaps for reasons along the lines Devitt suggests. It would be a mistake, however, to think that linguists in fact use intuitive beliefs concerning grammaticality as evidence for their hypotheses—even if they sometimes use the *label* 'grammaticality judgments' for what we label 'acceptability judgments.' Linguists may well—and may well be advised to—use their informed hunches as guides in their research. But they do not as a matter of fact justify their grammatical claims by adverting to whether some sentence *strikes them* as generated by the speaker's grammar. Nor would this seem a particularly strong consideration if they did.[3]

We therefore focus on the claim that linguists' *acceptability* judgments are more reliable indicators of syntactic reality. Devitt's own intentions, however, are unclear, owing to a conflation of acceptability and grammaticality. Devitt acknowledges the distinction ([2006b], pp. 101-2); but he argues that, in contexts controlled for non-grammatical sources of unacceptability (e.g., with minimal pairs), '"acceptable" acts a synonym for "grammatical"' and expresses the concept of grammaticality even if the subject has not lexicalized it (pp. 102, 110). This confuses what the term 'acceptability' means on that occasion (or, what concept it expresses) with what is the best explanation of the subject's judgment. The controls help narrow the space of candidate explanations; we see no reason to think they affect what the term means. Because Devitt conflates acceptability and grammaticality, there's no room to ask whether he's claiming that it's linguists' acceptability judgments or their (intuitive) beliefs concerning grammaticality that are more reliable indicators of syntactic reality. Having kept them apart, we examine the reading that renders the claim of interest.

On this reading, the claim is *not* that linguists' acceptability judgments tend to be *true*: to claim that would be to claim that linguists' acceptability judgments tend to be correct about *acceptability*. The claim rather is that among linguists there is a greater tendency for sentences they judge *acceptable* to be *grammatical* (or, more specifically, syntactically well-formed).[4] But reliability regarding acceptability is not irrelevant to our topic. Perhaps contrary to first impression, it is indeed possible to wrongly judge a sentence (un)acceptable: for instance, a subject may judge that 'The horse walked past the barn fell' is not a sentence that is naturally comprehensible or that she could use, but correct herself if primed with 'The horse that was walked past the barn fell'. And perhaps linguists are less likely to make such errors: years of training and experience may have improved their imaginative capacities in this domain. (Cf. Devitt [2006b], p. 111.) If linguists' acceptability judgments are in fact more reliable indicators of syntactic reality, this might be in part because their acceptability judgments are also more reliable indicators of acceptability. (Subjects' errors about acceptability, however, provide valuable data for theories of processing.)[5]

Two more matters before we turn to empirical results. First, a subject's reliability may vary with the conditions under which the judgments are made. If linguists are more susceptible to the effects of alcohol, they might not be more reliable *when in the vicinity of free beer*. More relevantly, what judgments subjects make might depend on whether they are given suitable instructions, or appropriately primed (as above), or presented sample strings in certain ways, etc. (See our discussion of Gordon and Hendrick [1997] below.) The question then is whether linguists are better subjects *under conditions C*. But what should conditions C include? This, again, depends not only on our goals as experimentalists, but also on our resources. For example, it is plausible that linguists would be better, though far from perfect, at screening out such non-grammatical sources of unacceptability as pragmatic oddity. But this advantage might diminish or vanish in conditions designed to neutralize such factors. Linguists would then not be better subjects, so far as reliability was concerned, if resources were invested to create such conditions. On the other hand, reliance on linguists' judgments (including one's own) might save one the effort. But this savings would not be worth it if, for example, there were a significant threat of theoreticians' bias or satiation to overcome. We note that, in our own experiment, judgments were elicited without pragmatic scene-setting and with only minimal manipulation of the conditions of judgment.

Second, how does one measure reliability? The obvious answer, in this case, is to compare acceptability judgments to the grammatical facts. But this is not what is done in previous studies such as those Devitt cites in support of his view. Perhaps this is in part because such a comparison requires that parties to the dispute antecedently agree about the grammatical facts—and it is possible that the relevant methodological disagreements (whether to rely on linguists' or non-linguists' judgments) would yield correlative theoretical disagreements concerning the grammatical facts. (Cf. Gordon and Hendrick [1997].) In any event, what the literature looks at instead is inter- and intra-group consistency: whether subjects in and across groups tend to make the *same* judgments. Intra-group consistency is necessary, but not sufficient for reliability. Inter-group *in*consistency suffices for at least one group's being unreliable. As will be apparent in the next section, more argument is needed to then draw conclusions about *which* group is more reliable. What is interesting about our result (reported in §4) is that, above a certain threshold of task-knowledge, we found no significant increase in correlation among subjects' judgments as a function of their knowledge of and expertise in theoretical linguistics. If that result were to hold up, there would be no basis for thinking linguists more reliable than non-linguists in the conditions we imposed.

3 Previous experiments

Before turning to our own results, we briefly discuss two papers—(Spencer [1973]) and (Gordon and Hendrick [1997])—that Devitt ([2006b], p. 111, fn. 25—cf. p. 115, fn. 36) cites as lending empirical support to the following claim:

> […] linguists have firm, and surely correct, intuitions about the acceptability of many sentences, and about some matters of co-reference, that the folk do not. ([2006b], p. 111)

It must be mentioned right away that neither paper draws the conclusion that linguists' judgments are more reliable. Indeed, the exact opposite is the case: having argued that linguists' and non-linguists' judgments *diverge*, both Spencer and Gordon and Hendrick conclude that *non-linguists'* judgments should be favored. Spencer ([1973], p. 90) in fact explicitly considers *but rejects* the Devittian contention that linguists' experience and knowledge might leave them better placed to detect the relevant properties (cf. Schütze [1996], p. 119, however, for critical discussion of her reasons). Gordon and Hendrick use the divergence they display to *reject* a syntactic theory based upon *linguists'* judgments in favor of one that accords with the judgments of their *non-linguist* subjects.

But presumably Devitt disagrees with the conclusions these authors draw from their data. He must mean only to direct our attention to the divergence they demonstrate. In fact, however, it's not even clear that these papers succeed in demonstrating that.

Spencer ([1973]) takes 150 sentences from several well-known papers in theoretical syntax and compares the judgment given to each sentence by the paper's author to the judgments of two groups of non-linguists: naïve non-linguists (students in an introductory psychology course), and non-naïve non-linguists (graduate students—some of whom were graduate students in linguistics—who had taken a course in generative grammar). She finds *inter alia* that the consensus judgments of both groups of non-linguists differed from the linguists' judgments on 59 sentences and that the consensus judgment of one group, but not both, differed from the linguists' on 14 more.

Schütze ([1996], pp. 115-9) canvasses a variety of concerns and replies: the various results were not tested statistically; the instructions (including their definition of grammaticality) are not supplied; none of the sentences are provided, though, judging from the source articles, many may be 'pragmatically very odd […] requir[ing] an usual context to sound acceptable,' something Spencer seems not to have controlled for (Schütze [1996], pp. 116—cf. Newmeyer [1983]); and so on. But the point that matters most to us is that Spencer compares the non-linguists' judgments to the judgment of a single linguist—the author of the paper in each case—rather than to a group of linguists. No conclusion therefore can be drawn about the intra-group consistency (or thus the reliability) of linguists on the whole.

The other study Devitt cites—(Gordon and Hendrick [1997])—examines non-linguists' judgments concerning various sentences relevant to the claims of Binding Theory. Crucially, they find that, unlike linguists, non-linguists do not accept co-reference between a name and a non-c-commanding, preceding pronoun, such as ***Her*** *brother visited* ***Sally*** *at college*.[6]

Gordon and Hendrick's results, however, also fail to give Devitt what he needs. First, they only look at a very small range of constructions. What is wanted—not for Gordon and Hendrick's purposes, but for Devitt's—is a comparison of judgments across a wide range of theoretically interesting constructions. Second, they don't test linguists as a group. They simply quote the judgments of a few linguists—Chomsky ([1981]), Postal ([1971]), and Higginbotham ([1980])—and note that it's not hard to find others who agree (Gordon and Hendrick [1997], pp. 338-9, fn. 6). But the sample is clearly not random: would a majority of syntacticians agree? Moreover, third, even the linguists they cite do not completely agree with one another, as Gordon and Hendrick (to their credit) observe. Postal, for example, questions a sentence—***His** father hates **John***—of which Higginbotham approves.[7] Finally, four, in the end they themselves partially explain away the divergence in terms of experimental design—specifically, by adverting to the effect of the surrounding sentences subjects are also asked to consider. Controlling for this factor, they find that a significant amount of the divergence vanishes. Indeed, we wouldn't be surprised if further pragmatic effects, beyond those they test for, might be in play and subject to control by priming. This might also explain why recent work seems to contravene their results: Kazanina et al. ([2007]), using methods in part based on Gordon and Hendricks', find that the relevant part of Binding Theory – Condition C – does indeed play an important role across a variety of syntactic environments.[8]

4 Our experiment

Let's turn now to our experiment. We tested 42 subjects: 7 Ph.D. linguists (syntacticians or linguists with substantial experience with research in syntax, age ranging from 25-64), 17 students with at least one class in generative syntax (ages 19-32), 11 students with no syntax background, but experience in other realms of cognitive science (ages 21-35), and 7 people with at least college-level education but no background in cognitive science (ages 24-61).[9] Notice that the age ranges are quite wide and that the age ranges (and spread) of those with the highest and lowest experience level are roughly equivalent.

Because we sought to uncover general differences among subjects' performance rather than the degree of acceptability for a particular construction, the stimuli consisted of 73 randomized sentences from an introductory linguistics class and textbook (Haegeman and Guéron [1999]), chosen to provide a mix of grammatical, ungrammatical, and questionable sentences. Each subject was provided with a questionnaire including instructions explaining the task. These instructions explicitly stated that the investigators were interested in speakers' intuitions of acceptability rather than prescriptive grammaticality. Acceptability was explained as follows:

> A sentence sounds good if you think you would or could say it under appropriate circumstances. By contrast, a sentence sounds bad if you think you would never say it under any circumstances.[10]

Subjects were then asked to read each sentence and rate it on a scale of 1-4 (1=perfect, 4=terrible). The questionnaire, including the instructions and a subset of sentences tested, can be found in the Appendix.

Each set of ratings provided by participants in a given group was compared using Spearman's rho (a non-parametric variant of Pearson's product moment correlation coefficient). The average correlation between subject in each group was computed, and p-values were estimated by Monte Carlo simulations over average correlation values. The results showed that the judgments of three groups—the linguists (identified as LOTS), the subjects with at least one class in theoretical syntax (SOME), and the subjects with no syntax experience, but with cognitive science background (LITTLE)—showed equally high intra-group average correlation values. However, the subjects with no experience in cognitive science of any kind (NONE) were *not* well correlated with one another: 30% of subject pairs were not significantly correlated, and the average correlation was significantly lower than the average correlations of the other three groups. Table 1 below shows these average intra-group correlations, and Figure 1 provides the Monte Carlo histograms.

[insert Table 1]

[insert Figure 1]

As for inter-group correlations (computed in a parallel fashion to the intra-group correlations), the same three groups (LOTS, SOME, and LITTLE) were highly correlated with one another and more correlated with each other than with the NONE group. Table 2 below shows these average inter-group correlations, and Figure 2 provides the Monte Carlo histograms.

[insert Table 2]

[insert Figure 2]

These data concern consistency, while Devitt's claim concerns reliability in the philosopher's sense. Consistency, recall, though necessary for reliability, does not in general *suffice* for reliability: subjects can be consistently *wrong*. In our case, however, the findings suggest that subjects with at least some relevant experience were much more reliable than subjects with no experience at all. For it's highly implausible that the groups with at least some relevant experience—including, the professional syntacticians—were consistently wrong (that is, consistently provided inaccurate indications of grammaticality), while the group of neophytes contained at least a higher number of accurate judgments. Much more plausible is that the greater consistency indicates greater reliability. The implausible alternative, in any event, would also disconfirm Devitt's claim.

Moreover, the results suggest that, contrary to what Devitt's claim would predict, a subject's knowledge of and experience in linguistics does not significantly affect performance, given a minimum level of relevant experience. The judgments of professional syntacticians were highly correlated with those of relative beginners in cognitive science.

But why is there a divide at all between group NONE subjects and the rest? It might be suggested that Devitt can offer an explanation that concedes our result while maintaining the core of his view. Perhaps subjects in group NONE have not passed a threshold for theoretical linguistic knowledge that the others have, and passing this threshold renders one so reliable an indicator of syntactic reality that further improvements in theoretical knowledge have little or no measurable effect. But this is implausible: why should taking a course in some other area of cognitive science enable one to cross such a significant threshold concerning linguistic knowledge?

It's likely, however, that subjects in groups LITTLE, SOME, and LOTS, unlike those in NONE, had either participated in these types of experiments before or had at least read or heard about studies that include these types of tasks. Such experience might be a significant factor in determining performance on this task. In particular, the lack of such experience might prevent the subject from interpreting the notion of acceptability as it is intended: for example, they might consider too narrow a range of counterfactual circumstances in determining whether they would use a sentence. In informal discussion, some subjects reported, for instance, that they rated a sentence unacceptable because they could think of a better way to say it or because it contained a specific lexical item they themselves would not use. This suggests that what group NONE subjects lack is task-specific knowledge, not expertise in syntax. One might test this hypothesis by investigating whether, e.g., some minimal training, more detailed instructions, or some control sentences might remove this effect.

What is noteworthy, however, is that no such interventions were needed to achieve high uniformity of judgment across minimally experienced subjects. Linguists, in this respect, do not make better subjects.

5 The context of Devitt's claim and his psychological model

We conclude by briefly mentioning the context in which Devitt enters his claim. This allows us to indicate some conclusions we do *not* intend our experiment to support and how our results raise a worry for Devitt's psychological model.

Devitt ([2006a], [2006b], chap. 7) argues that linguists' acceptability judgments are more reliable than non-linguists' in the course of replying to a possible argument for what he calls the Representational Thesis (RT). According to (RT):

> a speaker of a language stands in an unconscious or tacit propositional attitude to the rules or principles of the language which are represented in her language faculty. ([2006b], p. 273)

Devitt is concerned to rebut an abductive argument for (RT), according to which (RT) is part of the best explanation of why acceptability judgments are good evidence for linguistic theories. The argument proceeds roughly as follows:

> (1) Acceptability judgments are good evidence for linguistic theories.
> (2) This would be so if speakers derive these judgments via a causal/rational, deduction-like process from a representation of the rules of their language, so that the judgments are the 'voice of competence' (VC).
> (3) This explanation requires (RT).
> (4) There is no other, or better, explanation of (1).
> (5) So, (RT) is probably true.

Against this, Devitt argues, first, that the proposed explanation—(VC)—is problematic, and, second, that there is a better explanation. The better explanation is that:

> […] a normal competent speaker [….] uses herself as a guide to what the competent speaker *would do*. So she asks herself whether this expression is something she would say and what she would make of it if someone else

> said it. […] She does some central-processor reflection upon the datum to decide whether to apply her concept of grammaticality [sic] to the expression [….] Often these judgments will be immediate and unreflective enough to count as intuitions. Even when they do count, they are still laden with such background theory as she acquired in getting her concept of grammaticality. ([2006b], pp. 109-10)

It is to this explanation that Devitt adds the claim that, given the role of central-processor reflection in such judgments, and the greater knowledge and expertise of linguists, linguists should rely more on their own intuitions, as they are bound to be more reliable.

In providing evidence against this further claim of Devitt's, we do not intend to be offering even indirect support either for (VC) or (RT). Both (VC) and (RT) strike us as highly controversial. Indeed, although Devitt labels (VC) 'the standard linguistic answer' ([2006b], p. 96) to the question 'Why are speakers' intuitions evidence for linguistic theories?', it is also highly controversial whether either (VC) or (RT)—understood in Devitt's manner—is the considered view of any linguist, as Devitt ([2006b], p. 7) notes.[11]

We will not attempt a full assessment of Devitt's alternative psychological model. But our results suggest that he moves too quickly from the fact that acceptability judgments result from the interaction of various aspects of the mind/brain (in particular, are not the direct 'voice' of linguistic competence) to the claim that theoretical linguistic knowledge can significantly affect them. There are many positions intermediate between (VC), on the one hand, and the significant cognitive penetrability of acceptability judgments by acquired theoretical knowledge of linguistics, on the other.

Jennifer Culbertson
Department of Cognitive Science
Johns Hopkins University
3400 N. Charles St.
Baltimore, MD 21218
culbertson@cogsci.jhu.edu

Steven Gross
Department of Philosophy
Johns Hopkins University
3400 N. Charles St.
Baltimore, MD 21218
sgross11@jhu.edu

Funding

National Science Foundation (2006046914 to J. C., DGE 0549319 to Department of Cognitive Science, Johns Hopkins University).

Acknowledgments

For helpful discussion and advice, we thank Paul Bloom, Michael Devitt (especially), Géraldine Legendre, Amy Shelton, Paul Smolensky, Colin Wilson, members of Gross'

2007 graduate seminar, attendees of SPP 2008, participants in the Dubrovnik Philosophy of Linguistics workshops, and the referees.

References

Chomsky, N. [1965]: *Aspects of the Theory of Syntax*, Cambridge, MA: MIT Press.

Chomsky, N. [1981]: *Lectures on Government and Binding*, Dordrecht: Foris Publications.

Devitt, M. [2006a]: 'Intuitions in Linguistics', *British Journal for the Philosophy of Science* **57**, pp. 481-513.

Devitt, M. [2006b]: *Ignorance of Language*, Oxford: Oxford University Press.

Evans, G. [1980]: 'Pronouns', *Linguistic Inquiry* **11**, pp. 337-62.

Ferreira, F. [2005]: 'Psycholinguistics, Formal Grammars, and Cognitive Science', *The Linguistic Review* **22**. pp. 365-80.

Gordon, P., and Hendrick, R. [1997]: 'Intuitive Knowledge of Linguistic Co-Reference', *Cognition* **62**, pp. 325-70.

Haegeman, L., and Guéron, J. [1999]: *English Grammar: A Generative Perspective*, Oxford: Blackwell Publishing.

Higginbotham, J. [1980]: 'Pronouns and Bound Variables', *Linguistic Inquiry* **11**, pp. 679-708.

Kazanina, N., Lau, E., Lieberman, M., Yoshida, M., and Phillips, C. [2007]: 'The Effect of Syntactic Constraints on the Processing of Backwards Anaphora', *Journal of Memory and Language* **56**, pp. 384-409.

Marantz, A. [2005]: ‘Generative Linguistics within the Cognitive Neuroscience of Language’, *The Linguistic Review* **22**, pp. 429-45.

Newmeyer, F. [1983]: *Grammatical Theory: Its Limits and Its Possibilities*, Chicago: University of Chicago Press.

Phillips, C. [forthcoming]: ‘Should We Impeach Armchair Linguistics?’, in S. Iwasaki (*ed.*), *Japanese/Korean Linguistics*, **17**, Palo Alto: CSLI Publications.

Postal, P. [1971]. *Cross-Over Phenomena*, New York: Holt, Rinehart and Winston.

Schütze, C. [1996]: *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*, Chicago: University of Chicago Press.

Snow, C., and Meijer, G. [1977]: ‘On the Secondary Nature of Syntactic Intuitions’, in S. Greenbaum (*ed.*), 1977, *Acceptability in Language*, The Hague: Mouton, pp. 163-77.

Spencer, N. [1973]: ‘Differences between Linguists and Nonlinguists in Intuitions of Grammaticality-Acceptability’, *Journal of Psycholinguistic Research* **2**, pp. 83-98.

Appendix

INSTRUCTIONS:

In this task, you will be asked to judge whether individual sentences sound good or not TO YOU. A sentence sounds good if you think you would or could say it under appropriate circumstances. By contrast, a sentence sounds bad if you think you would never say it under any circumstances. Here, we are interested in your linguistic intuitions, not in the rules of "proper English" that you might have been taught in school. For example, consider sentence (1):

(1) Mary never goes nowhere.

You may have been taught that sentences with double negatives as in (1) are not acceptable in "proper English". However, under certain circumstances you may actually produce this sentence yourself. If this is the case then you should NOT judge (1) as bad. Basically, we would like you to ignore any language rules you might have been taught and focus strictly on whether you think you would or could say the sentence under appropriate circumstances.

With these things in mind, please read each of the following questions and rate them according to the scale provided below. In order to indicate your response, circle the number corresponding to the rating you wish to give the sentence. If you do not understand the scale or have questions about any of the sentences please ask the experimenter. You will have 15 minutes to complete this portion of the study, so you should NOT spend a lot of time trying to figure out what linguistic rules may be violated in each sentence. We are interested in your IMMEDIATE reactions to the sentences.

Rating Scale:

1 PERFECT: This sentence sounds perfect, you would use it under appropriate circumstances.

2 OKAY: This sentence is not completely perfect, but it still pretty good, and you might say it under appropriate circumstances.

3 AWKWARD: This sentence sounds strange, you doubt you would ever say it.

4 TERRIBLE: This sentence sound terrible, you would never say it under any circumstances.

SENTENCE **RATING**

1. I promise that on no account during the holidays will I write a paper. 1 2 3 4

2. All hell is about to break loose. 1 2 3 4

3. It is important he be on time. 1 2 3 4

…

71. Sharon asked what he think that he will eat for lunch. 1 2 3 4

72. I told her the reason why I wondered whether she would invite him. 1 2 3 4

73. I promise that on no account will I write a paper during the holidays. 1 2 3 4

[Note to type-setter: The material in the Appendix above required us to include some extra formatting. We would like the final product to look something like that, even if you need to alter our formatting. Thank you.]

[Type-setter: Tables and figure captions below. Figures supplied in a separate file.]

[table 1, to be inserted above:]

Table 1. Intra-group correlations

	Avg r^2
LOTS	0.482
SOME	0.428
LITTLE	0.495
NONE	0.201*

*significant at α=0.05 (Bonferroni corrected)

[figure 1's caption, to be inserted above:]

Figure 1. Results of the Monte Carlo simulations for intra-group correlation (red lines represent actual average r^2 values for each group)

[table 2, to be inserted above:]

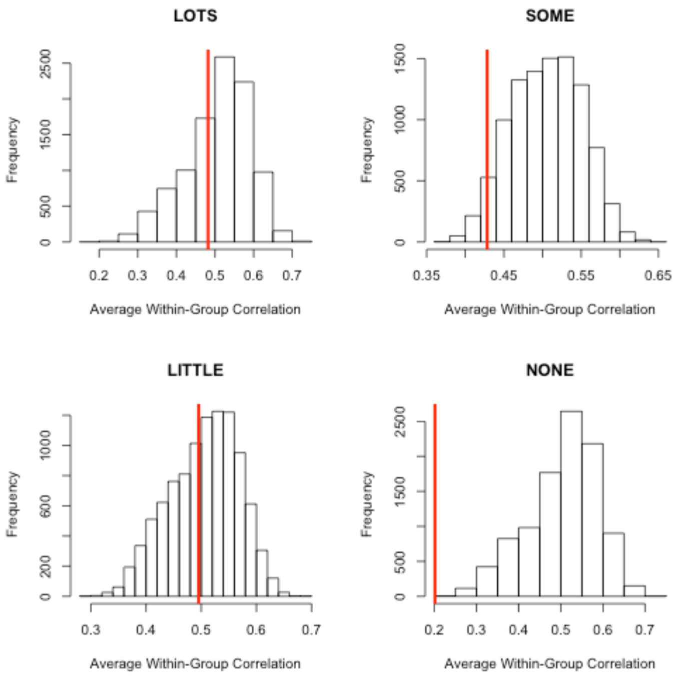
Table 2. Inter-group correlations

	Avg r^2		
	LITTLE	**SOME**	**LOTS**
NONE	0.302*	0.281*	0.283*
LITTLE		0.464	0.470
SOME			0.450

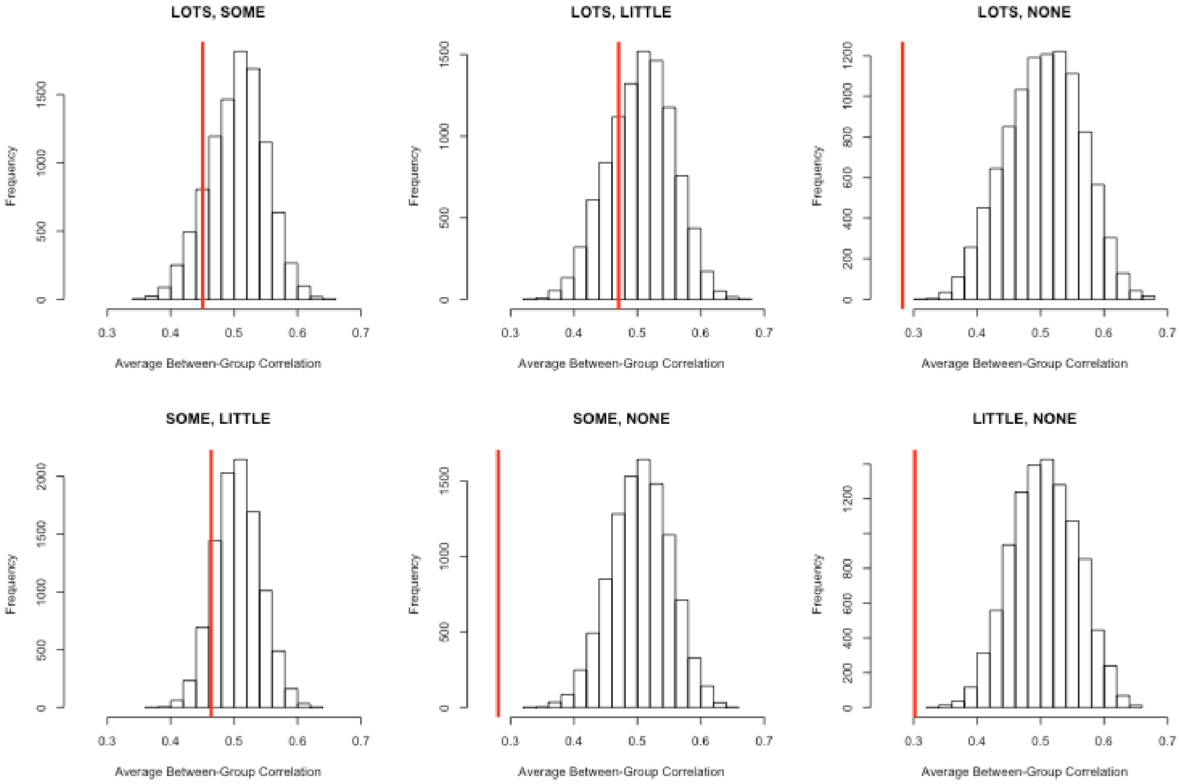
*significant at α=0.05 (Bonferroni corrected)

[figure 2's caption, to be inserted above:]

Figure 2. Results of the Monte Carlo simulations for inter-group correlation (red lines represent actual average r^2 values for each group)



[type-setter: above is Figure 1]



[type-setter: above is Figure 2]

[Type-setter: Below are endnotes, to be converted to footnotes for printed version:]

[1] Among the various other sources of data that might be brought to bear on beliefs concerning grammaticality are reaction times, developmental patterns, acquired deficits, dissociations, ERP and fMRI studies, computational results, corpus studies, and elicited utterances. Cf. (Devitt [2006b], pp. 98-9). Beliefs concerning grammaticality must also mesh with relevant theorizing about related domains, as the appeal to memory limitations illustrates.

Devitt ([2006b], chap. 2) has a view concerning what beliefs concerning grammaticality—or, the claims of grammatical theory—are *about* that differs from the view of many linguists. Devitt holds that they are not about language users' I-languages (an aspect of their mind-brains), but rather about properties of external linguistic items that relevant mind-brain processes must respect. This dispute, like those mentioned in §5 below, is orthogonal to our topic.

Though we place great stress on the acceptability/grammaticality distinction, we would not go so far as to claim that no linguist has ever been unclear on the matter. Newmeyer ([1983]), for instance, suggests that some linguists' notation—their use of stars and question marks—has been ambiguous between ungrammaticality and unacceptability. On the other hand, Schütze ([1996], p. 45, fn. 23) replies that 'the authors of works reviewed here, like most current authors, intend the latter interpretation.'

[2] For comprehensive discussion, see (Schütze [1996]). For more recent contributions, see, e.g., (Ferreira [2005]; Marantz [2005]; and Phillips [forthcoming]).

[3] On some fine-grained conceptions of concept-individuation, it's not only unsurprising, but trivial that linguists' beliefs concerning grammaticality are more reliable than non-linguists'; for, on these views, most non-linguists do not so much as *possess* the theoretical concept of grammaticality in the first place and so can't even form such judgments. We doubt, however, that Devitt would accept such a conception of concept-individuation. We take no position on this ourselves.

[4] This could provide a reader—who brackets considerations of charity—grounds for ascribing to Devitt the uninteresting claim that linguists' beliefs concerning sentences' *grammatically* are more reliable indicators of syntactic reality; for, generally, when philosophers talk of a class of beliefs' reliability, they have in mind their tendency to be accurate concerning *what they represent*, not their reliability as indicators of some other target.

[5] Schütze ([1996], p. 61) discusses studies showing that 'what forms people actually use [diverges from] what they claim they use.' But he appears not to see how this divergence bears on his earlier discussion of whether acceptability judgments are objective. There he says that 'they cannot be verified or resolved by observation or calculation.' (p. 52)

[6] Gordon and Hendrick explain in their instructions to subjects that bold indicates co-reference. In addition to the divergence noted above, they also find divergence in judgments concerning some sentences with co-referring proper names, such as ***John*** *said that* ***John*** *would win*.

[7] Another paper they cite—(Evans [1980])—likewise disputes others' judgments concerning co-referring proper name cases like the one in fn. 6 above.

[8] Perhaps a study more favorable to Devitt's claim is (Snow and Meijer [1977]). They find significantly higher intra-group consistency among linguists than among non-linguists. On the other hand, they also find a high correlation between linguists' and non-linguists' judgments, contrary to Spencer.

[9] Devitt ([2006b], pp. 109, 115) speculates that subjects with *no* formal education will perform markedly differently on linguistic judgment tasks. We did not test such subjects. However, Scribner and Cole ([1981]—cf. Schütze [1996], pp. 122-4), reporting field work among the Vai people of Liberia, found that literacy and schooling had little effect on linguistic judgment tasks, though there was some effect on subjects' *explanations* of their judgments.

[10] Alternatively, the instructions might have explained acceptability in terms of whether one would or could understand *someone else's* utterance of the sentence. This might have blocked judgments based on personal preferences (see below). But it might have admitted many otherwise excluded sentences: for example, we can often readily understand what a novice 2nd-language learner intends to communicate in uttering a string that is ungrammatical in our language, such as 'She seems sleeping.' Conjoining the alternative instruction with the actual instruction (cf. Devitt [2006b], p. 109) might have blocked such cases, but without affecting judgments based on personal preference.

[11] Interestingly, Schütze ([1996], p. 6), in attempting to discover whether anyone has explicitly maintained a position like (VC), is reduced to offering the indirect evidence that Lasnik, Chomsky, and others have felt the need to argue against it. But cf. (Devitt [2006b], pp. 96 and 97, fn. 3).

For spirited debate about this and other aspects of Devitt’s views, see the special issues devoted to the Philosophy of Linguistics in the 2006 and 2008 volumes of the *Croatian Journal of Philosophy*.