3

Pair-list answers in naïve speakers and professional linguists

Asya Achimova, Peter Staroverov, Viviane Déprez and Julien Musolino

3.1 Introduction

Informally collected grammaticality judgments have probably been the most widely used kind of data in generative linguistics. Although such judgments can be pretty robust (Sprouse & Almeida 2010, 2012a, Sprouse, C. T. Schütze & Almeida 2013), disagreements among professional linguists in their judgments of particular sentences have doubtlessly arisen. In such cases, collecting judgments in a formal experimental setting has proven useful (C. T. Schütze 1996, 2009, Kawahara 2011, C. T. Schütze & Sprouse 2014). Moreover, professional linguists have sometimes been reported to differ in their judgments from naïve speakers of the same language (Spencer 1973, Gordon & Hendrick 1997, Culbertson & Gross 2009, Dąbrowska 2010, Gibson & Fedorenko 2010, 2013). This latter kind of disagreement, if upheld, could be particularly worrisome as they carry implications that linguists could have concerned themselves with phenomena that are largely idiosyncratic to their group (as some authors conclude, see e.g. Gibson & Fedorenko 2010, 2013).

However, before such negative conclusions can be drawn, we need to gain a better understanding of the nature of the difference in grammaticality judgments between professional linguists and naïve speakers (C. T. Schütze & Sprouse 2014). This paper compares naïve speakers and linguists in an experimental study of semantic acceptability judgments for scopally ambiguous sentences. We show that, as a group, naïve speakers and professional linguists give similar judgments. However, it also turned out that some naïve speakers (about 30% in our initial study) were likely to accept scopal interpretations previously judged unacceptable by most linguists. A further investigation of this difference in two follow-up studies, showed first that naïve speakers are more susceptible to task effects than linguists, and second, that they may be more likely to unconsciously accommodate a sentence to a correct one via lexical substitution. When these effects are appropriately controlled for, naïve speakers' judgments become closer to those of linguists. Consequently, this study argues that while naïve speakers and professional linguists have the same grammatical competence, the judgments of the former are more likely than those of the latter to be affected by performance factors (Spencer 1973, Newmeyer 1983, 2007, Devitt 2006). Furthermore, such performance factors may be especially strong when judgments concern subtle semantic distinctions that are bound to particular situations, rather than more straightforward grammaticality ones.

The paper is structured as follows. In Section 3.1.1 we review the literature concerned with naïve speakers' vs. professional linguists' judgments. Section 3.1.2 briefly introduces the linguistic phenomenon used in our study. Section 3.2 describes our Experiment 1, which compares naïve speakers and professional linguists in their judgments of semantic acceptability. Section 3.3 describes two follow-up studies designed to further investigate the nature of the qualitative differences that surfaced between linguists and naïve speakers. Section 3.4 presents the cumulative discussion of the results and our conclusions.

3.1.1 Grammaticality Judgments and the Judgment Providers

In a recent review article C. T. Schütze & Sprouse (2014: 27) cite the choice of a population of judgment providers as "one of the most contentious aspects of judgment data". Indeed there is a growing literature documenting the differences between professional linguists and naïve speakers in their judgments (Spencer 1973, Gordon & Hendrick 1997, Culbertson & Gross 2009, Dąbrowska 2010). In most of these studies the reported differences between the two groups are qualitative rather than quantitative. While overall naïve speakers as a group behave statistically very similarly to professional linguists, the patterns of variation by subject diverge. The present study reveals a similar pattern with respect to semantic acceptability judgments.

Two kinds of explanations have been offered for the observed differences between naïve speakers and linguists. First, it has been suggested that linguists could be subconsciously biased towards giving judgments that confirm their own theoretical beliefs (Edelman & Christiansen 2003, Ferreira 2005, Wasow & J. Arnold 2005, Gibson & Fedorenko 2010, 2013). Dąbrowska (2010) addressed this concern in a study of how professional linguists rate island effects. Island effects represent important empirical phenomena extensively investigated within the generative grammar framework. At the same time, the grammatical nature of island effects has been questioned both among generative linguists and among functional linguists alike. In a study that compared island violations ratings by generative linguists with those of functional linguists, Dąbrowska (2010) showed that the generative linguists turned out to rate island violations as more acceptable than the functional linguists did, as if the former were biased *against* their own theoretical conclusions.

Second, differences between linguists and naïve speakers have been attributed to a heightened sensitivity by the former to relevant differences, or a greater capacity to ignore certain irrelevant factors that affect the overall sentence well-formedness (Spencer 1973, Newmeyer 1983, 2007, Devitt 2006). It was observed that linguists can potentially more easily abstract away from individual lexical items, the plausibility of scenarios they are assessing, the complexity of sentences — the factors introducing confounds that can interfere with acceptability judgments in naïve speakers. In short, it would seem that linguists understand better what the task is. Although the linguists' heightened sensitivity can be difficult to prove, there is some existing experimental evidence that provide suggestive support for this type of explanation. Culbertson & Gross (2009)sought to investigate the role of expertise on judgments by looking at how consistent speakers of each group turn out to be. Defining judgment reliability as consistency in responses in different circumstances, regardless of accuracy, they tested professional linguists with substantial experience in syntax, students with at least 1 course worth of experience in generative syntax, and a group of naïve subjects with no experience in cognitive science. A comparison of students who had experience in generative syntax and of another student group who only had experience in other domains of cognitive science was intended to help revealing whether the amount of task-specific knowledge affects the quality of judgments. Subjects were asked to evaluate sentences from a syntax textbook (Haegeman & Guéron 1999).The analysis shows that speakers with some task-specific knowledge were more consistent in their responses as a group (showed less variability), and hence were more reliable. The authors acknowledge the fact that consistency does not necessarily imply reliability in terms of actual reflection of true syntactic processes. However, they suggest, it seems rather implausible that a group of naïve speakers could have had more accurate judgments than speakers with some level of expertise for no particular reason.

Interestingly, the amount of experience in linguistics did not affect the consistency of judgments in any substantial way. Culbertson & Gross (2009) suggest that the uniformity of judgments is achieved through minimal task specific knowledge, and does not reflect knowledge of linguistic theory. In other words, the divide would lie between speakers who have never performed linguistic judgment tasks as opposed to those who have had some experience participating in such tasks (see also Devitt 2010, Gross & Culbertson 2011 for further discussion). As we will see, the results of the present study go in the same direction. They suggest that linguists are indeed more sensitive to subtle semantic differences than naïve participants, but also show that certain manipulations of the judgment task can make it easier for naïve speakers to detect the relevant linguistic distinctions (see also Fanselow 2007, Grewendorf 2007, Haider 2007).

A final important issue, that we only partially address here, concerns potential distinctions between judgments that are reported in the linguistic literature and judgments by linguists or naïve speakers that are elicited in controlled experiments (Gibson & Fedorenko 2010, 2013, Sprouse & Almeida 2012a). Concerned with this issue, Gibson & Fedorenko (2013) examined a number of case studies; one of these involves superiority violations in multiple *wh*-questions. According to the Superiority condition (Chomsky 1973), in a well formed muliple *wh*-question (direct or embedded) that contains both a subject and an object question, it is the *wh*-subject phrase, i.e. the hierarchically highest phrase that must front and the *wh*-object, i.e the structurally lowest phrase, that must remain in its original position, as in (1). Cases in which the reverse occurs lead to unacceptability, as in (2) as the Superiority condition is violated.

(1) Peter knows who bought what.

(2) *Peter knows what did who buy.

(3) Peter knows what did who buy where.

However, according to Bolinger (1978) and Kayne (1983), the addition of third *wh*-phrase, such as *where* in (3), is reported to improve the acceptability of such superiority violation. Gibson & Fedorenko (2013) put this claim to an experimental test using embedded questions. They found, contra existing claims in the theoretical literature, that naïve speakers found no differences between sentences like (2) and (3) and proceeded to conclude that naïve speakers data collected in experimental conditions had to be used to avoid possible bias effects that could lead theoretical generalizations astray.

Conclusions of Gibson & Fedorenko (2010, 2013) were later challenged in a number of papers (Culicover & Jackendoff 2010, Sprouse & Almeida 2010, 2012b). Sprouse & Almeida, in particular, questioned the logic of their conclusions arguing that differences found between judgments reported in the literature and data elicited from naïve speakers do not constitute evidence that the latter type of data is the only reliable one. Existing large-scale controlled studies of syntactic judgments have indeed confirmed that the majority of informal judgments reported both in textbooks (Sprouse & Almeida 2012a) and in linguistic journals (Sprouse, C. T. Schütze & Almeida 2013) are reliably replicated experimentally with naïve participants.

The present study compares three groups of speakers judging the asymmetric availability of pair-list answers in identical experimental settings: undergraduate students, Ph.D. candidates in linguistics, and professional linguists with a Ph.D. We show that, overall, judgment patterns are consistent across groups, although individual patterns of variation can emerge. Importantly, we also show that judgments across different groups of speakers can be collectively similar even for sentences whose acceptability has been debated in the literature, as our brief review section of the literature on the relevant linguistic phenomenon attests.

3.1.2 Subject-object Asymmetries in Wh*-/quantifier Interactions*

In their ability to variably license so called *pair-list answers*, or PLAs for short, questions with quantifiers are a prime example of the linguistic complexity that characterizes the interactions of scope bearing elements. Observing that PLAs are only available for questions in which a universal quantifier occurs in a subject position, as in (4), but not for questions in which the quantifier occurs in an object position, as in (5), May (1985) can outscope *wh*-elements that are fronted above them only under syntactically limited circumstances.

(4) Which boy did every girl kiss?
Mary kissed John, Sue kissed Nick, and Helen kissed Michael.

(5) Which girl kissed every boy?
*Mary kissed John, Sue kissed Nick, and Helen kissed Michael.

A number of distinct accounts for the rather famous contrast in (4–5) long regarded as a standard case of the subject-object asymmetry have been proposed (May 1985, Chierchia 1993, Beghelli 1997, Agüero-Bautista 2001). While all existing accounts converge in predicting the asymmetry given in (4–5), the various proposed theories diverge in the consequent set of varying empirical predictions they make in regards to modifications of this basic paradigm. Although our experiments focus on the judgments that are common to all accounts, it is important to note that various data points remain controversial in the literature, offering evidence that the judgments data surrounding this particular research question are far from trivial.

The original account in May (1985) treats the asymmetry in (4–5) as a consequence of a general syntactic principle: in (5), the object quantifier fails to outscope the question term, because its LF movement would violate the Path Containment Condition (Pesetsky 1982) by crossing the movement path of the *wh*-item. As shown by Beghelli (1997), however, there are lexical differences among quantifiers in regards to the basic asymmetry: strongly distributive quantifiers like *each* appear to be able to outscope a question term even when they occur in object positions (see also Williams 1988, Szabolcsi 1997a, Agüero-Bautista 2001) as witnessed by their ability to have PLAs in questions like (6). Beghelli takes this to show that *each*, unlike *every*, can raise to the specifier of a designated projection Dist(ributive)P, located higher than IP, from which it can bind the variables introduced by the *wh*-phrase (Beghelli 1997).

(6) Which girl kissed each boy? PLA ok.
$[_{CP}$ Which girl$_j$ $[_{DistP}$ each boy$_i$ $[_{IP}$ t$_j$ [kiss $[_{NP}$ t$_i]_i]]]]$

For him, on the other hand, weakly distributive quantifiers like *every* that are lexically underspecified for distributivity cannot raise to DistP.

Focusing on the nature of question terms in contrast, Chierchia (1993) suggests that PLAs may be available with an object quantifier in questions with a semantically plural *wh*-term like *who*, but not with a strictly singular question term like *which* in (5). Chierchia further proposes to analyze restrictions on PLAs as a consequence of general binding conditions, and more specifically, as resulting from Weak Crossover effects that prevent the binding of a pronominal variable by a non-c-commanding quantifier. Notably, such effects are suspended with semantically plural pronouns, thus explaining why

PLA could be unrestricted with plural questions terms. Similar judgments for *who*-questions are reported in Agüero-Bautista (2001), for whom the ability for a *wh*-phrase to give rise to PLAs depends on restrictions that govern the reconstruction of a question term below the interacting quantifier according to the presuppositional status of a *wh*-phrase, and not its plurality.

Table 3.1 summarizes the empirical predictions of the accounts briefly reviewed above.[1]

Subject questions	May (1985)	Beghelli (1997)	Chierchia (1993)	Agüero-Bautista (2001)
Who kissed every girl?	–	–	+	+
Which boy kissed every girl?	–	–	–	–
Which boy kissed each girl?	–	+		+

Table 3.1: Availability of pair-list answers for subject questions with object quantifiers.

As discussed in details in Achimova, Déprez & Musolino (2013) and as shown by Table 3.1, all these accounts agree on the unavailability of PLAs for questions like (5) (*which* interacting with *every*) and also manifest a relative consensus on availability of PLAs for questions like (6) (*which* interacting with *each*).However when it comes to the potentially plurality of *who* and the use of *which* in plural contexts, the predictions diverge. The availability of PLAs to questions with quantifiers thus presents an ideal testing ground for assessing the differences between linguists and naïve speakers. The reported judgments in this case involve a subtle and complex semantic phenomenon, and manifest both partial convergence and debated discrepancies in the literature.

3.2 Experiment I: Professionals vs. Naive Speakers

3.2.1 Methods

Design The experiment was designed to test whether the predicted subject-object asymmetry exemplified in (4–5) above can be verified for three groups of

[1] Plus signs indicate that a PLA is predicted to be possible and minus signs — unavailable.

speakers differing in their level of linguistic training. We kept the question/answer pairs as close as possible to those discussed in the literature. Crossing the factors resulted in a 2 × 2 × 2 × 3 design: 2 (*quantifier position:* subject vs. object) × 2 (*answer type:* single vs. pair-list) × 2 (*wh-type: who* vs. *which*) × 3 (undergraduate students, Ph.D. candidates in linguistics, professional linguists with a Ph.D.).

Participants The undergraduate group contained 33 psychology students who received course credit for their participation. We also tested 32 Ph.D. candidates in linguistics, and 28 professional linguists holding a Ph.D., all native speakers of English. We recruited our subjects through the Linguist List. Professional linguists were also asked whether they were familiar with the literature on *wh-*/quantifier interaction and pair-list answers. The level of familiarity with the topic did not affect the ratings to target items in the experiment ($p = 0.55$).

Materials and procedure Each trial consisted of a questions/answer pair. The task was to determine whether that particular answer was a *possible* answer to the relevant question on a 1–7 scale, where 1 was 'definitely no' and 7 'definitely yes'. A sample question is given in (7).

(7) Which driver took everybody home last night?
Tom took Ms. Franko, Bob took Ms. Dombovski, and Jack took Mr. Perkins.

Participants were asked to rate 32 critical items and 60 control/filler statements which included questions with clearly acceptable or unacceptable answers, as well as questions with pragmatically odd answers. The experiment started with the presentation of three trial stimuli. Participants then took the main test that lasted between 15–20 minutes.

3.2.2 Experiment I: Results

The analysis was performed using cumulative link mixed models (R package 'ordinal'). We first fit a model with ratings as a dependent variable and type of answer as an independent variable, random effects include random intercepts for subjects and items and random slopes for subjects. As expected, single answers received higher ratings (mean = 6.8 on a 7-point scale) than PLAs (mean = 5) ($\beta = 4.4, SE = 0.513, p < 0.01$). Single answers serve as control, showing that subjects had no problems dealing with questions containing

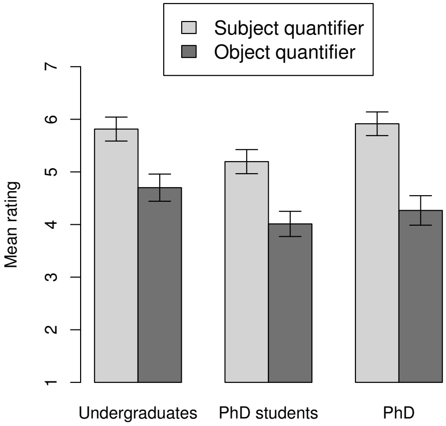

Figure 3.1: Subject/object asymmetry for different groups of speakers

universal quantifiers overall. From now on, our analysis focuses solely on PLAs since it is about their distribution that conflicting claims are made.

The analysis confirmed a significant effect of quantifier position: PLAs to questions with subject quantifiers received higher ratings, than PLAs to questions with object quantifiers as predicted by all approaches ($\beta = 2.49, SE = 0.36, p < 0.01$). Professional linguists did not differ from either naive subjects ($\beta = 0.42, SE = 0.57, p = 0.46$), or Ph.D. students in linguistics ($\beta = -0.39, SE = 0.57, p = 0.49$) with regards to this type of question/ answer pair. These results confirm the literature findings of the subject-object asymmetry in the distribution of PLAs for all the tested populations.

We now turn to a more detailed analysis of the responses. Figure 3.2 shows the distribution of ratings assigned by the speakers to PLAs in questions with object quantifiers. Further analysis revealed that among naïve speakers at least 30% assigned a rating of 6 or 7 to such question-PLAs pairs, in contrast to the predicted unavailability of PLAs in such cases (May 1985, Beghelli 1997). However, the number of speakers showing no subject-object asymmetry appears to diminish with expertise. It is smallest for professional linguists.

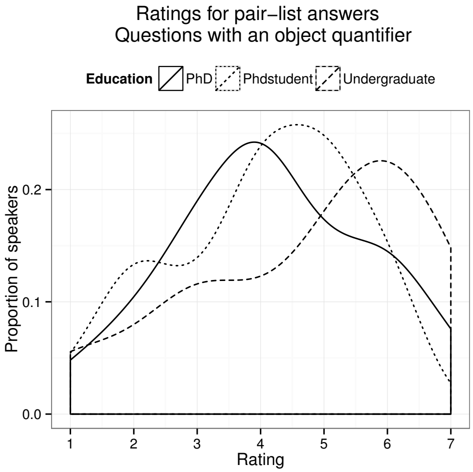

Figure 3.2: Distribution of ratings (averages across 8 items of a given type)

3.2.3 Experiment I: Discussion

The results of Experiment I are generally in line with what is typically observed in the literature (see 3.1.1). On the one hand, professional linguists, Ph.D. candidates, and naïve participants as a group give very similar results, and all groups confirm the presence of the subject-object asymmetry. On the other hand, the patterns of variation in judgments are different between the three groups. While very few professional linguists with a Ph.D. judged PLAs to object-quantifier questions to be possible, more Ph.D. students in linguistics did so (i.e. consistently rating these 6 or 7), and even more naïve participants (at least 30%).

Could this pattern of judgments indicate that 30% of the naïve participants have a different grammar (being then perhaps less likely to become linguists)? We contend that this is rather unlikely, and suggest instead that naïve participants could be more amenable to ignoring certain confounds. For one thing, naïve participants may be more willing to accommodate than linguists. When accepting PLAs to object-quantifier questions with *every,* undergraduate students may unconsciously accommodate the distributivity of *every*, making it, in relevant respects, more similar to the quantifier *each*. Recall from Section 3.1.2 that strongly distributive quantifiers like *each* are known to escape the subject-object asymmetry observed with the pseudo-distributive ones like *every* (Beghelli 1997). If some of our naïve participants subconsciously accommodated *every* to *each,* this would predict a higher acceptability ranking for object-quantifier questions[2]. In Experiment 2, we show that this subconscious lexical accommodation can be avoided when participants are asked to judge sentences with *every* alongside sentences with *each,* thus increasing their awareness of the contrast.

Another possible reason why relatively many naïve participants seem to accept the supposedly ungrammatical PLAs may have to do with the set up of the task. Naïve speakers lack the experience of producing acceptability judgments, and therefore may be more susceptible to noise that could be introduced by the choice of fillers and control items in an particular experiment. We address this concern in Experiment 3.

3.3 Follow-up Experiments

The experimental methods for both Experiment 2 and Experiment 3 were essentially the same as for Experiment I, although only naïve speaker participants took part in the follow-up studies. In Experiment 2 participants were asked to judge answers to questions with that vary the type of quantifier *every* vs. *each* in addition to its position. As a consequence, it is plausible to suppose, that their awareness of the contrast between these two quantifiers was sharpened, making them less likely to accommodate *every* to *each.* We see in Figure 3.3 that this resulted in a shift of the mode of ratings for *every* object-quantifier questions as compared to the results of Experiment 1, suggesting that the contrast between *every* vs. *each* is indeed relevant to naïve speakers' judgments.

In Experiment 3, we asked naïve speaker participants to perform the same task but the number of items per condition was increased up to 20, and a binary yes/no judgment was used instead of a scale. The set of controls was also modified: instead of using pragmatically incoherent answers as unacceptable items (8), questions with downward entailing quantifiers such as *nobody*, *most*, and *few* were used, resulting in pairs like (9).

[2] Interestingly in this regards, naïve speakers behave not unlike preschoolers for whom as Achimova, Syrett, et al. (submitted) show, the distributivity contrast between *each* and *every* is inexistent.

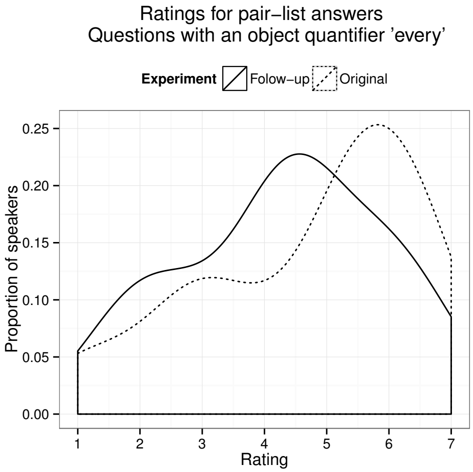

Figure 3.3: Naïve participants

(8) Did you read every book on the list?
Yes, I read 3 out of 8

(9) Who did nobody see?
Mary didn't see John, Sue didn't see Nick, and Helen didn't see Mike.

The results of Experiment 3 are summarized in Figure 3.4.

If displaying the expected subject-object asymmetry, participants are predicted to accept PLAs with subject-quantifier but not with object-quantifiers questions. Hence, data points should cluster in the upper left part for each of the right and left graphs (high rating/acceptance rate for subject-quantifier questions, and low rating/acceptance rate for object-quantifier questions). In the original experiment (left graph) we see that at least 30% of speakers show *similarly* high acceptance for PLAs in both the subject- and the object-quantifier condition. This is not true however for the follow-up (yes/no) experiment, where participants show behavior in line with theoretical predictions: participants clearly rejected PLAs to questions with object quantifiers.

Because several parameters were modified in this follow-up experiment, it is possible that all of them contributed in sharpening the subject-object asymmetry for naïve speaker participants. Note, however, that several studies

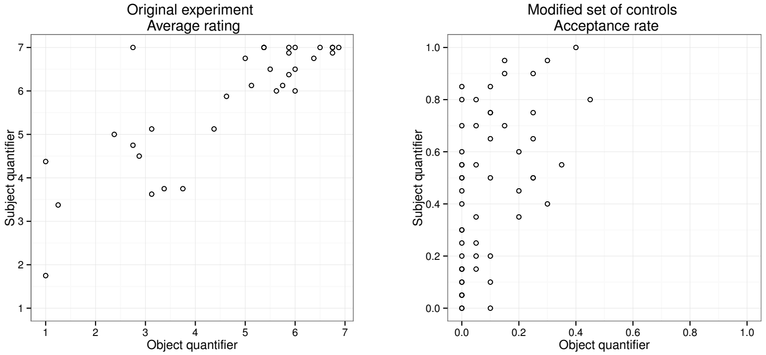

Figure 3.4: The effect of control items in an experiment

have shown that using a scale vs. a binary a yes/no judgment task produced essentially similar results (Bader & Häussler 2010, Kawahara 2011). Increasing the number of tested items should likewise have little effect on judgment quality; though possibly help in producing a cleaner quantitative picture of the responses. Thus the factor that is most likely to be responsible for the effect observed in Figure 3.4 must come from using a different set of controls/fillers. In this follow-up experiment, we used controls/fillers that more closely matched the type of violation expected in the critical items. We conjecture that in being asked to compare sentences with different quantifier types, the sensitivity to the task might have been increased. Conversely, it is possible that the set of controls used in Experiment I created an overly strong impression of deviance that belittled the comparatively more mild deviance of object-quantifier PLAs for naïve participants. In sum, it would appear that the type of comparison class items used as controls in a judgment task is of importance in sharpening the attention of naïve speakers to pertinent contrasts.

3.4 Discussion

Pair-wise comparisons of professional linguists, linguistics students and naïve speakers did not reveal an effect of expertise on the ratings in Experiment I. Thus our experimental results indicate that speakers of all three groups essentially patterned alike: they manifested a clear subject-object asymmetry in their rating of PLA availability, and variability in judgments was present for all three groups of speakers for the controversial object-quantifier questions like (5), but not for the subject-quantifier questions like (4).

We observed that judgments tended to get closer to those reported in the literature (rating a PLA to an object-quantifier question lower) as expertise increases, yet the analysis revealed no statistical differences between professional linguists and naïve speakers. This implies that data from experts and naïve speakers can be a reliable source of acceptability judgments. This result is advantageous because naïve speaker subjects are often easier to access as a population, and when useful, experiments can be performed with larger numbers of speakers.

Our results also offer some insight into the differences that are here observed between linguists and naïve speaker participants. In line with the sensitivity hypothesis outlined in 3.1.1, we argued that linguists are more able to abstract away from certain performance factors that can act as confounds. In the case at hand, it appears that there were at least two potential sources of such confounds. First, Experiment 1 only tested questions with *every*, but the availability of very similar questions with *each* for which the PLAs are acceptable has apparently led some naïve participants to accommodate and rate PLAs higher than expected from the theoretical literature. Second, the nature of the fillers and controls used in Experiment 1 may have made it more likely for naïve participants to apply the accommodation strategy, because unacceptable controls were of a rather different nature than the critical items, and clearly very degraded, being not just grammatical deviant, but also discursively incoherent. The results of Experiments 2 and 3 suggest that such confounds can be addressed by making naïve speaker participants more aware of important lexical contrasts and by choosing control items that set up more appropriate linguistic contrasts. When these factors are adequately controlled for, the variation within the group of naïve speaker participants becomes very similar to that observed with more expert linguists in Experiment I. We conclude that although both naïve speaker participants and linguists can give very consistent judgments, experiments with the former group should be carefully designed to address the potential effects of scale adjustment and accommodation. We further submit that the type of controls used in linguistic experiments should also be detailed as their nature may well be of central importance in influencing the judgment of non-expert naïve speakers.