

This printout has been approved by me, the author. Any mistakes in this printout will not be fixed by the publisher. Here is my signature and the date _____.

Is Magnitude Estimation Worth the Trouble?

Shin Fukuda¹, Grant Goodall², Dan Michel², and Henry Beecher²

¹University of Hawai'i at Mānoa and ²University of California, San Diego

1. Introduction

In sentence acceptability experiments, subjects are typically asked to indicate their response to sentences in one of three ways. In a yes/no forced choice task, subjects simply indicate whether or not the sentence sounds good. In an n-point numerical scale (or Likert scale) task, the extremes of a numerical scale are defined (e.g. 1 = “sounds very bad” and 5 = “sounds very good”) and subjects choose a number on the scale that reflects their overall response to the sentence. Finally, in a magnitude estimation (ME) task, subjects compare experimental sentences to a reference sentence. This reference sentence is associated with a number (or subjects may choose this number on their own). Subjects are instructed to rate the experimental sentences in relation to the number given to the reference sentence. If the experimental sentence sounds twice as good as the reference sentence, for instance, subjects are to multiply the reference sentence’s score by 2; if it sounds half as good, they should divide it in half, etc.

Each of these response methods has potential advantages and disadvantages. The yes/no method has the virtue of being very easy for subjects to understand, but it is often thought to be relatively coarse-grained and to require large numbers of subjects to detect fine differences. The n-point numerical scale arguably yields finer-grained results, while still being easy for subjects, but there is no guarantee that the chosen scale will allow as many distinctions in acceptability as subjects actually perceive or that subjects will treat the distance between any two adjacent points on the scale as being the same (e.g. in a 5-point scale, subjects might treat the difference between 1 and 2 as being larger or smaller than that between 3 and 4). ME is clearly not easy for subjects to understand, in that it is an unfamiliar task that requires some mathematical sophistication, but it could reasonably be expected to overcome the two disadvantages seen for the n-point scale. Subjects are able to make as many distinctions as they want, and since they are explicitly asked to make ratio judgments (i.e. how many times better or worse the experimental sentence is compared to the reference sentence), one would expect less uncertainty about the nature of the results.

In this study, we submit these three response methods to a critical examination by comparing the results obtained by each in three otherwise identical experiments. In section 2, we review the previous literature on these methods, and in section 3, we present the set of experiments that constitute the core of our contribution, concluding that some of the claimed advantages of ME do not appear to be empirically supported. We devote some attention to differences among the three methods that emerged in our results in section 4 and we explore some other results of interest in our experiments in section 5. Section 6 presents conclusions and implications for the working syntactician.

2. Previous studies

For the reasons sketched in the introduction, ME is often taken to be the “gold standard” of response methods in sentence acceptability experiments, and it is assumed to produce a fundamentally different type of data and provide insights that are not possible with other methods (e.g., Bard et al.

* Thanks to members of the audience at WCCFL and to members of the UCSD Experimental Syntax Lab for their very helpful and stimulating comments. Thanks are due as well to the audience at the 2010 LSA meeting, where we presented a preliminary version of this paper, to Jon Sprouse for many helpful suggestions and discussions, and to Sara Cantor for invaluable assistance. All errors remain our own.

1996, Cowart 1997, Featherston 2005). Recently, however, this gold standard status of ME has been scrutinized from two different directions. First, some recent studies have questioned whether the assumptions behind ME indeed hold. For instance, one of the crucial assumptions behind the alleged superiority of ME is that it produces ratio-based judgments. However, Sprouse (2011) shows that participants in ME experiments of sentence acceptability do not seem to be making ratio-based judgments. Second, ME has also been examined in terms of the empirical results produced. Wescott and Fanselow (2008, 2011) had two groups of participants judge a set of stimuli that consists of German sentences with object scrambling with different case marking (dative vs. accusative). The participants judged the same set of stimuli twice with a two-week interval in-between, either with (a) a forced choice and then a numerical scale task or (b) a forced choice and then an ME task. Bader and Häussler (2010) also had two groups of participants judge the same stimuli with two tasks: forced-choice (speeded grammaticality judgment) and ME, both in a single experiment. Unlike Wescott and Fanselow, they examined three different phenomena in German with different degrees of expected acceptability contrasts: (i) object scrambling with different case marking, a relatively clear contrast, (ii) German equivalents of *be* and *get* passives with accusative and dative objects, a relatively subtle contrast and (iii) permutations of orders among three verbs in a verb cluster with a full range of acceptability, from clearly acceptable to completely unacceptable. Both the Wescott and Fanselow and the Bader and Häussler studies conclude that the results of ME experiments are no more informative than the results of forced choice tasks or numerical scale tasks. Wescott and Fanselow in particular argue that ME results contain a greater amount of spurious variance.

The present study is similar to these latter papers in that its main aim is to compare empirically the results obtained from collecting sentence acceptability judgments with different methods. Yet it differs from these previous studies in two important ways. First, we had three different groups of participants judge the same set of stimuli, but each group used a different method: a forced choice task, a numerical scale task or ME. While having the same participants judge the same stimuli with different methods, as both Wescott and Fanselow and Bader and Häussler did, avoids potential issues that might arise from comparing the results from two different groups of participants, the fact that participants rated the same stimuli twice with different methods in these studies raises a different set of potential problems, such as the possible effects of the first method on the second or changes in judgments upon repeated exposure to the same sentence types. Second, our stimuli consist of phenomena with a wider range of contrasts than the stimuli of Wescott and Fanselow and Bader and Häussler. We examined three different syntactic phenomena from English with different degrees of expected acceptability contrasts: (i) presence/absence of Subject-Auxiliary verb inversion (henceforth Subj-Aux inversion) in *wh*-questions, an extremely clear contrast, (ii) the *that*-trace effect, a relatively clear contrast and (iii) subextraction from embedded subject, object, and *wh*-subjects, a relatively subtle contrast. In addition, we included an extra factor in (i) by alternating the type of subjects between (a) 2nd-person pronouns, (b) 3rd-person pronouns and (c) lexical DPs (e.g. *the man*). This addition was inspired by the fact that in similar environments in Spanish, acceptability varies in extremely subtle (though statistically significant) ways depending on the type of subject (Goodall 2010). If something similar occurs in English, including this factor in the experiment will allow us to see to what extent each method succeeds in capturing these extremely subtle contrasts, which are likely ultimately due to extra-grammatical considerations.

3. Experiment

3.1 Subjects

A total of 108 undergraduate students at the University of California, San Diego, all self-identified native speakers of English, participated for course credit. They were randomly assigned to one of the three different methods: a yes/no forced-choice task (henceforth y/n), a 5-point numerical scale task and ME. There were thus 36 participants in each group.

3.2 Methods

An identical set of stimuli was presented to the participants in all three methods. Participants first

received brief instructions about the assigned method and then had a practice session (4 items) prior to the actual experiment. Participants were instructed not to analyze the sentences, but to give their first reaction by rating how good or bad the sentences sounded to them. For ME, a *wh*-question with marginal acceptability (*What do you wonder whether Mary bought?*) was used as the reference sentence and it was given a score of 100.

3.3 Materials

As briefly discussed in section 2, the experiment consisted of three subexperiments. Subexperiment 1 used a 2 x 3 design to test the effects of the presence and absence of Subj-Aux inversion in English *wh*-questions with three subject types: 2nd-person pronouns, 3rd-person pronouns and lexical DPs. Subexperiment 2 had a 2 x 3 design, crossing THAT (presence vs. absence of *that*) and EXTRACTION (embedded subject extraction, embedded object extraction and no extraction). This allows us to test for the *that*-trace effect, the well-known phenomenon in which extraction from embedded subject position in the presence of *that* is degraded. Subexperiment 3 examined the effect of the argument type/position of a DP on subextraction from it. It had a 2 x 3 design, crossing EXTRACTION (presence vs. absence of extraction) and ARGUMENT (embedded subject vs. embedded object vs. embedded *wh*-subject).¹ Six lexicalizations for each of the 22 conditions were constructed and distributed among 6 lists using a Latin Square procedure. Each subexperiment served as fillers to the others, and 5 additional filler items were added to each list, resulting in 27 total items. The lists were pseudo-randomized in two ways, yielding 12 lists in total. Examples of the stimuli are provided below:

(1) Subexperiment 1: INVERSION x SUBJECT TYPE:

- a. What will you/he/the man watch on Thursday? (inversion x three subject types)
- b. What you/he/the man watch on Thursday? (no inversion x three subject types)

(2) Subexperiment 2: THAT x ARGUMENT:

- a. Who do you feel that ___ insulted Pat at the theater? (*that* x subject extraction)
- b. Who do you feel that Pat insulted ___ at the theater? (*that* x object extraction)
- c. Do you feel that Pat insulted Mary at the theater? (*that* x no extraction)
- d. Who do you feel ___ insulted Pat at the theater? (no *that* x subject extraction)
- e. Who do you feel Pat insulted ___ at the theater? (no *that* x object extraction)
- f. Do you feel Pat insulted Mary at the theater? (no *that* x no extraction)

(3) Subexperiment 3: EXTRACTION x ARGUMENT:

- a. What do you think pictures of ___ will be on the website? (extraction x embedded subject)
- b. What do you think the website will post pictures of ___? (extraction x embedded object)
- c. What do you wonder which pictures of ___ will be on the website? (extraction x embedded *wh*-subj)
- d. Do you think pictures of the new car will be on the website? (no extraction)
- e. Do you think the website will post pictures of the new car? (no extraction)
- f. Do you wonder which pictures of the new car will be on the website? (no extraction)

As discussed in section 2, each of the subexperiments presented syntactic phenomena from English with varying ranges of acceptability, from very clear contrasts (presence and absence of Subj-Aux inversion) to very subtle, potentially extra-grammatical contrasts (different types of subjects in presumably grammatical and ungrammatical sentences).

3.4 Statistical Analysis

The raw ratings obtained from the 5-point task and ME were converted to z-scores in order to

¹ The experiment contained a fourth subexperiment examining the interaction of inversion in *wh*-questions and the argument vs. adjunct status of the *wh*-phrase (2 x 2 design). For reasons of space, this subexperiment is not discussed in this paper.

facilitate comparison across methods. These results were then analyzed using linear mixed-effects models with the factors identified in section 3.3 as fixed factors and participants and items as random factors. Multiple pair-wise comparisons were also conducted to isolate the effect of particular factors. The binary judgments obtained from the y/n method were analyzed using logistic mixed-effects models using the same fixed factors and random factors.

3.5 Results

3.5.1 Subexperiment 1: Subj-Aux Inversion

Subexperiment 1 examined the effect of Subj-Aux inversion. Since our focus here is on the effects of the presence/absence of inversion, we present the results in Figure 1 with all three subject types collapsed.

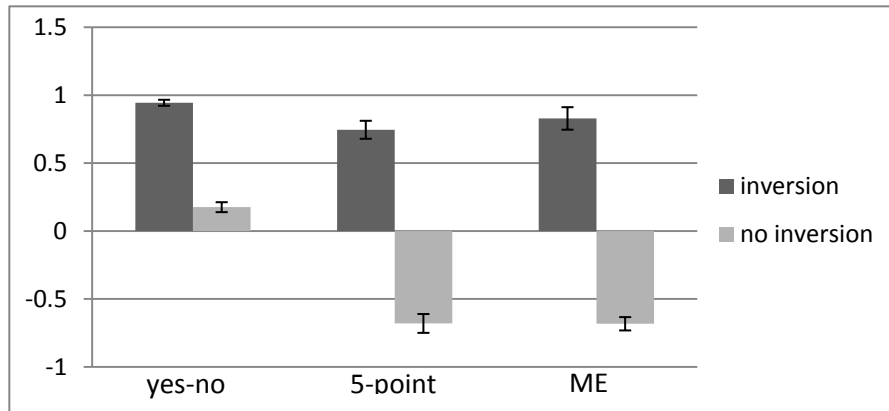


Figure 1. Subj-Aux Inversion with the three subject types collapsed

To facilitate visual inspection, Figure 1 presents the results from the three methods side-by-side. It should be noted, however, that for the y/n method, the y-axis represents mean scores where 1 = yes and 0 = no. For the 5-point and ME tasks, on the other hand, the y-axis represents the z-score, where a positive value shows that the given mean is higher than the overall mean and a negative value shows that it is lower. As Figure 1 shows, *wh*-questions that lacked Subj-Aux inversion were significantly less acceptable than their counterparts with Subj-Aux inversion in all three methods (y/n: $p < .001$, 5-point: $p = .0001$, ME: $p = .0001$).

3.5.2 Subexperiment 2: That-trace effects

Subexperiment 2 examined extraction from embedded subject and object position with and without *that*. No-extraction conditions were also included in this subexperiment, but those will be omitted here for reasons of space. Figure 2 presents the results from the three methods:

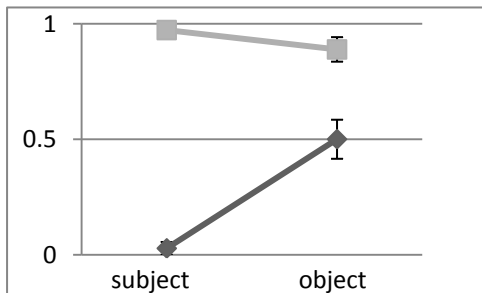


Figure 2a. Extraction with yes/no

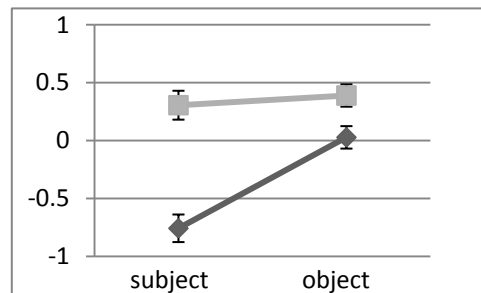


Figure 2b. Extraction with 5-point scale

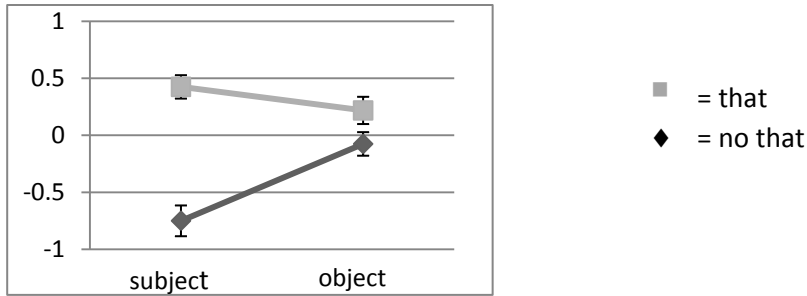


Figure 2c. Extraction with ME

The results from all three methods indicate that (i) the interaction between THAT (presence vs. absence of *that*) and ARGUMENT (subject vs. object) was significant (y/n: $p = .0014$, 5-point: $p = .0058$, ME: $p = .0014$), (ii) ARGUMENT was a significant predictor of acceptability of *wh*-questions with THAT (y/n: $p = .0013$, 5-point: $p = .0002$, ME: $p = .0016$) but was not a significant predictor of acceptability of *wh*-questions without THAT (y/n: $p = .989$, 5-point: $p = .662$, ME: $p = .314$) and (iii) THAT was a significant predictor of acceptability of *wh*-questions with subject extraction (y/n: $p < .0001$, 5-point: $p = .0001$, ME: $p = .0002$). A difference among the three methods is observed with *wh*-questions with object extraction. While the results from the y/n and 5-point methods indicate that THAT was a significant factor with *wh*-questions with object extraction (y/n: $p = .001$, 5-point: $p = .0454$), the results from ME indicated that it was not ($p = .123$).

3.5.3 Subexperiment 3: Subextractions

Subexperiment 3 examined subextraction from embedded objects, embedded subjects, and embedded *wh*-subjects, with their non-extraction counterparts used as controls. Figure 3 presents the results from the three methods:

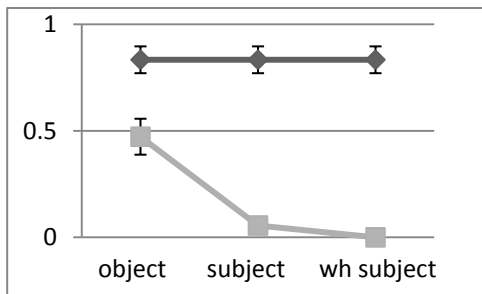


Figure 3a. Subextraction with yes/no

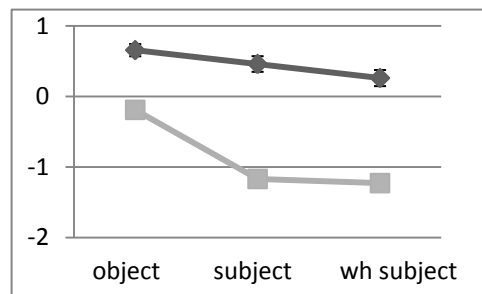


Figure 3b. Subextraction with 5-point scale

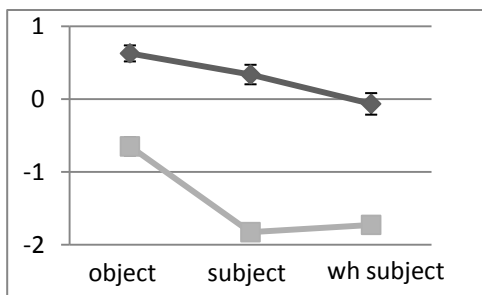


Figure 3c. Subextraction with ME

The results of the y/n method and the 5-point task indicate that the interaction between EXTRACTION (presence vs. absence of extraction) and ARGUMENT (embedded subject vs. embedded object vs. embedded *wh*-subject) was significant (y/n: $p = .0218$, 5-point: $p = .0027$). However, it was only marginally significant with ME ($p = .0582$). Another difference among the three methods was observed with sentences with the no EXTRACTION condition. With these, ARGUMENT was a significant factor only in the 5-point task (y/n: $p = .929$, 5-point: $p = .041$, ME: $p = .14$). Yet the three methods gave virtually the same results with respect to the rest of the factors. All three methods indicated that ARGUMENT was a significant factor within sentences with the EXTRACTION condition (y/n: $p = .0052$, 5-point: $p = .0001$, ME: $p = .0002$). Pair-wise comparisons among ARGUMENT factors revealed that, within the EXTRACTION condition, subjects and objects (y/n: $p = .0052$, 5-point: $p = .0001$, ME: $p = .0036$) as well as *wh*-subjects and objects (y/n: $p < .001^2$, 5-point: $p = .0001$, ME: $p = .0001$) were significantly different from each other. In contrast, none of the methods indicated that there was a significant difference between subjects and *wh*-subjects (y/n: $p = .997$, 5-point: $p = .7504$, ME: $p = .2554$).

3.5.4 Subexperiment 1 revisited: Subject Types with/without Inversion

As discussed earlier, Subexperiment 1 examined not only the effect of the presence/absence of Subj-Aux inversion, but also the interaction between Subj-Aux inversion and three different subject types: 2nd-person pronouns, 3rd-person pronouns and lexical DPs. Our aim with this additional factor was to see (i) whether these three subject types have an effect on the acceptability of *wh*-questions without Subj-Aux inversion and (ii) if so, whether the three methods differ in their ability to capture these very subtle, possibly extra-grammatical contrasts. Figure 4 presents the results from the three methods:

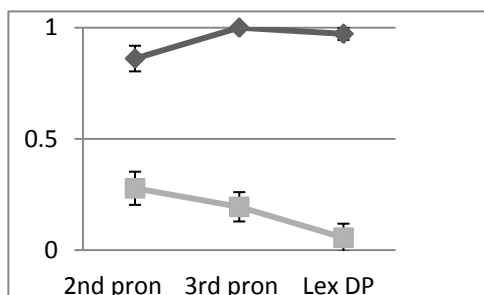


Figure 4a. Subject types with yes/no

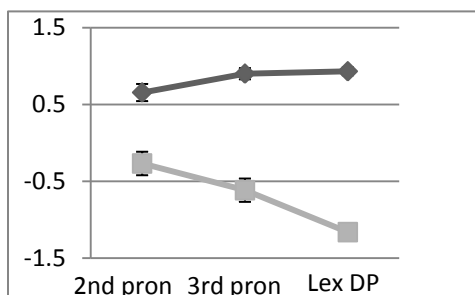


Figure 4b. Subject types with 5-point scale

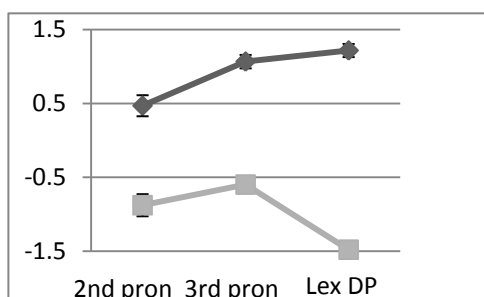


Figure 4c. Subject types with ME

◆ = inversion
 ■ = no inversion

² The contrast in subextraction between objects and *wh*-subjects with y/n task came out as not significant in the logistic mixed effects model. This is presumably due to the known property of these models that they sometimes fail with smaller samples for a contrast that is too clear (Myers 2009). Given that the mean for the *wh*-subject subextraction condition was 0, with no variance, the measure of significance reported here is based instead on a Wilcoxon rank sum test.

None of the three methods found SUBJECT TYPE to be a significant predictor of acceptability for *wh*-questions with INVERSION (although it came close with ME (y/n: $p = .456$, 5-point: $p = .108$, ME: $p = .062$)). In *wh*-questions without INVERSION, however, all three methods found SUBJECT TYPE to be significant (y/n: $p = .0367$, 5-point: $p = .0001$, ME: $p = .0028$), suggesting that contrasts among subject types exist and that all three methods are able to capture them, despite their subtlety. Moreover, all three methods found 2nd-person pronoun subjects to be significantly more acceptable than lexical subjects in the -INVERSION case (y/n: $p = .0287$, 5-point: $p = .001$, ME: $p = .008$). Additionally, the 5-point and ME tasks both found a significant contrast in this case between 3rd-person and lexical subjects (5-point: $p = .0014$, ME: $p = .0242$, cf. y/n: $p = .14$), but none of the methods found a contrast between the two pronominal subject types with no INVERSION (y/n: $p = .4672$, 5-point: $p = .15$, ME: $p = .2886$). In sum, all three methods converge on SUBJECT TYPE being a significant factor in *wh*-questions without INVERSION and on there being a significant difference between 2nd-person and lexical subjects in particular.

3.5.5 Summary of results

For an impressive array of cases, then, the three methods examined here capture virtually the same contrasts, finding significant differences in acceptability both in very clear and in much more subtle contexts. These include the difference between inversion and non-inversion structures in *wh*-questions (subexperiment 1), the *that*-trace effect (subexperiment 2), subextraction from objects, subjects, and *wh*-subjects (subexperiment 3), and the effect of subject type on *wh*-questions without inversion (subexperiment 1 again). The general conclusion is that for the types of contrasts that are of interest to syntacticians, all three methods appear to be sufficiently sensitive to capture very fine gradations in acceptability.

4. Some remarks on differences

We have seen that the three methods examined here provide strikingly consistent results across a wide range of contrast types, from very clear to very subtle. Nonetheless, there are some interesting ways in which they differ, as we have seen. First, the yes/no method differs from the other two in one contrast that appears to be a straightforward case of lack of sensitivity. In both the n-point scale and ME methods, a *wh*-question without inversion is significantly better with a 3rd-person pronominal subject than with a lexical subject. The yes/no method does show a numerical advantage for the pronominal subject, but this does not reach significance. The explanation for this difference is probably very simple: the difference in acceptability between these two sentence types is extremely small and the yes/no method has not been able to capture it here, though perhaps it would with a larger sample size.³

Second, the yes/no and 5-point methods show a significant difference between the presence and absence of *that* in object *wh*-questions, whereas in ME, these two sentence types are statistically indistinguishable. Given our present knowledge, it is hard to know what to conclude from this. On the one hand, it may be that there is a real contrast between these two sentence types (see Cowart (1997, 2003) for some relevant data) and what we see here is a case where ME is actually a less sensitive measure than the other two (see Wescott and Fanselow (2011) for related discussion). On the other hand, it may be that there is no true contrast in acceptability between the two sentence types and that the yes/no and 5-point methods are simply giving us a type I error here (see Kim and Goodall (in press), for example, who do not find this contrast with a much larger sample size). The data from our study do not allow us to choose between these possibilities with any confidence.

³ See Sprouse and Almeida (submitted), however, who present the results of experiments that show that the yes/no method consistently required smaller numbers of participants to reach acceptability contrasts than ME did when testing the same sets of acceptability contrasts.

5. Other results of interest

In addition to the methodological implications addressed up to this point, the experiment also contains some empirical findings that are of interest in their own right because of how they relate to ongoing concerns in syntactic theory. We discuss three such findings here, dealing with the *that*-trace effect, subextraction from DPs, and subject type in inversion.

5.1 *That-trace effect*

One unequivocal result of the present study is that there is a very robust *that*-trace effect. That is, with all three methods, a statistically significant difference was found between extraction of a subject in the presence of *that* and two other sentence types: extraction of an object in the presence of *that* and extraction of a subject in the absence of *that*. This result is of interest because it has become common to claim that the *that*-trace effect is highly variable, with sizeable numbers of speakers not showing the effect (see, e.g., Sobin 1987, Rizzi and Shlonsky 2007). In the present study, we find no evidence for this view (see also Cowart 2003). Though we cannot exclude the possibility that there might be populations of native English speakers without a *that*-trace effect (see Kim and Goodall (in press) for such a case among non-native populations), this is no more true for this effect than it is for any of the syntactic phenomena examined here. This in turn suggests that the *that*-trace effect is not accidental in English and should follow in a deterministic way from other properties of the language.

5.2 *Subextraction from DPs*

All three methods in the present study found a significant difference in subextraction from a subject vs. object DP. This confirms the widely accepted view that subject DPs are islands (e.g., Chomsky 1973). More interesting is the fact that no method found a significant difference between subextraction from a subject vs. a fronted *wh*-phrase. This is perhaps a surprising result, since many have suggested that *wh*-phrases allow subextraction more easily than subjects do (e.g., Chomsky 1986, Lasnik & Saito 1991, Boškovič 2002).

There are two possible conclusions that one may draw from this latter result. First, it may be that none of the methods used here are sensitive enough (given the number of subjects and the number of tokens of each sentence type) to capture the difference in acceptability between these two types of subextraction. Alternatively, it may be that there simply is no difference, as has been claimed in some of the recent literature (e.g., Chomsky 2008). If true, this allows for an analysis in which subextraction is allowed out of *in situ* phrases, such as objects, but not out of moved phrases, such as fronted *wh*-phrases and subjects in English (e.g., Stepanov 2007).

5.3 *Subject type in inversion*

All three methods in this study found that *wh*-questions without inversion that have 2nd-person subjects (*you*) are rated significantly higher than those that have lexical subjects (e.g., *the man*). This is an unexpected result, since both sentence types are ruled out by standard analyses and there is no obvious syntactic basis for a distinction between them in English. As mentioned earlier, Goodall (2010) finds a similar distinction in Spanish, but he shows that given the structure of *wh*-questions in that language and the way that they behave cross-dialectally, such a distinction is to be expected. Whether such an account may be extended to English remains an open question for future research.

6. Conclusion

Let us return now to the main empirical result of this paper: the three methods examined here show an overwhelming consistency in their results, regardless of whether the contrasts being explored are very clear or very subtle. For all three subexperiments, the same major results were obtained in the great majority of cases no matter what response method was used. This finding is of course particularly meaningful for the yes/no and 5-point tasks, since they have long been thought to be less

able than ME to capture fine gradations in acceptability. If they are approximately equal to ME in this respect, as our experiment suggests, this then removes one of the main purported advantages of ME. If we also consider the well-known disadvantages of ME, such as the fact that it is more difficult for experimenters to implement and requires more mathematical sophistication on the part of experiment participants, the yes/no and n-point tasks then appear to be more reasonable choices for most sentence judgment experiments.

To the extent that the difficulties associated with ME have discouraged many syntacticians from adopting experimental techniques, that roadblock may now be safely removed. Simple, straightforward response methods such as the yes/no forced-choice task and the n-point numerical scale task appear to be sufficient to capture the types of contrasts in sentence acceptability judgments that are of interest to syntacticians.

References

- Bard, Ellen Gurman, Dan Robertson and Antonella Sorace (1996). Magnitude estimation of linguistic acceptability. *Language* 72: 32-68.
- Bader, Markus. and Jana Häussler (2010). Toward a model of grammatical judgments. *Journal of Linguistics* 46: 273-330.
- Bošković, Željiko (2002). A-movement and the EPP. *Syntax* 5:167–218.
- Chomsky, Noam (1973). Conditions on Transformations. In Stephen R. Anderson and Paul Kiparsky (eds.), *Festschrift for Morris Halle*, New York: Holt, Rinehart & Winston.
- Chomsky, Noam (1986). *Barriers*. Linguistic Inquiry Monograph 13. Cambridge, MA: The MIT Press.
- Chomsky, Noam (2008). On Phases. In Robert Freidin, Carlos P. Otero, María Luisa Zubizarreta (eds.), *Foundational Issues in Linguistic Theory. Essays in Honor of Jean-Roger Vergnaud*. Cambridge, MA: The MIT Press. pp. 133–166.
- Cowart, Wayne (1997). *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks: Sage Publications.
- Cowart, Wayne (2003). Detecting syntactic dialects: The *that*-trace phenomenon. Paper presented at the 39th meeting of the Chicago Linguistic Society, April, Chicago.
- Featherston, Sam (2005). Magnitude estimation and what it can do for your syntax: Some WH-constructions in German. *Lingua* 115: 1525-50.
- Goodall, Grant (2010). Experimenting with *wh*-movement in Spanish. In Karlos Arregi, Zsuzsanna Fagyal, Silvina A. Montrul and Annie Tremblay (eds.), *Romance Linguistics 2008: Interactions in Romance*. 233-48. John Benjamins: Amsterdam.
- Kim, Boyoung and Grant Goodall (in press). Age-related effects on constraints on *wh*-movement. In Darren Tanner and Julia Herschensohn (eds.), *Proceedings of the 11th Generative Approaches to Second Language Acquisition Conference*. Sommerville, MA: Cascadilla Proceedings Project.
- Lasnik, Howard and Mamoru Saito (1992). *Move α : Conditions on its application and output*. Cambridge, Mass.: MIT Press.
- Myers, James (2009). The design and analysis of small-scale syntactic judgment experiments. *Lingua* 119: 425-44.
- Rizzi, Luigi and Uri Shlonsky (2007). “Strategies of Subject Extraction”, in H.-M.Gärtner and U. Sauerland (eds.) *Interfaces + Recursion = Language? Chomsky's Minimalism and the View from Syntax-Semantics*. 115-16. Berlin: Mouton de Gruyter
- Sobin, Nicholas (1987). The variable status of Comp-trace phenomena. *Natural Language and Linguistic Theory* 6: 445-501.
- Sprouse, Jon (2011). A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language* 87: 274-288.
- Sprouse, Jon and Diogo Almeida (submitted). *Power in acceptability judgment experiments*.
- Stepanov, Arthur (2007). The end of CED? Minimalist and extraction domains. *Syntax* 10: 80-126.
- Wescott, Thomas and Gisbert Fanselow (2008). Variance and informativity in different measures of linguistic acceptability. *Proceedings of the West Coast Conference on Formal Linguistics (WCCFL)* 27: 431-439.
- Wescott, Thomas and Gisbert Fanselow (2011). On the informativity of different measures of linguistic acceptability. *Language* 87: 249-73.