Is Magnitude Estimation Worth the Trouble?

keywords: experimental syntax, wh-movement, magnitude estimation

There are three response methods typically used in sentence acceptability judgment experiments: categorical (yes/no), n-point scale (Likert), and magnitude estimation (ME). Of these, ME is widely regarded to be superior (e.g., Bard et al. 1996, Cowart 1997). Nonetheless, the unfamiliar nature and cumbersome procedure of ME has meant that it has not gained broad acceptance among syntacticians, which has in turn significantly limited the use of experimental methods in syntax in general (Myers 2009). In addition, the putative superiority of ME has recently been challenged. For instance, it has been shown that the role of the modulus in ME experiments is at best unclear (Sprouse 2008, 2010) and ME can be less informative than the other two methods (Weskott and Fanselow 2008). The current study provides further evidence that experimental syntax should not be limited to ME; categorical and n-point response methods can capture crucial contrasts just as well and have the virtue of being easier to perform.

To examine possible differences in sensitivity to grammatical contrasts, the three response methods were used to collect acceptability judgments with wh-questions in three domains: (1) Subject- auxiliary inversion, (2) that-trace effect, and (3) extraction out of subject DPs, object DPs, and fronted wh-phrases (all shown here with standard judgments from the literature).

(1) a. What will you watch on Thursday?
   b. *What you will watch on Thursday?

(2) a. *Who do you feel that ___ insulted Pat at the theater?
   b. Who do you feel ___ insulted Pat at the theater?

(3) a. *What do you think [pictures of ___] will be on the website?
   b. What do you think the website will post [pictures of ___]?
   c. *What do you wonder [which pictures of ___] will be on the website?

These contrasts range from very clear (as in (1)), to very subtle (as in (3)). In addition, sensitivity of inversion to the type of subject DP (2nd-person vs. 3rd-person vs. lexical) was probed.

(4) a. What will you/he/the boy watch on Thursday?
   b. *What you/he/the boy will watch on Thursday?

A preference for a 2nd-person subject in cases like (4b) has been noted for Spanish (e.g. Goodall 2010).

Materials: The stimuli presented to subjects consisted of 3 subexperiments: (I) 6 conditions for subject- auxiliary inversion (type of subject x presence/absence of inversion), (II) 4 conditions for that-trace (extraction of subject vs. object and presence/absence of that), and (III) 6 conditions for extraction from DPs/wh (subject/object/wh x presence/absence of extraction). 6 lexicalizations were created for each of the conditions and distributed among 6 lists using a Latin Square design. There were 36 subjects (native-speaker undergraduates) for each of the three methods (108 subjects total); each subject judged the stimuli using only one method (randomly assigned). A 5-point scale was used for the n-point scale.

Results: Results are summarized below. The three methods find significant contrasts in almost exactly the same set of cases, ranging from the relatively clear contrasts in (1) and (2) to the much more subtle contrasts in (3) and (4). There are also findings that are interesting in their own right: No method finds a contrast between extraction out of subject vs. wh-phrase (3b vs. c), and all three find a contrast in non-inversion between 2nd-person vs. lexical subject (4b). This striking consistency across methods is blemished only by two (arguably minor) ways in which the categorical method is unlike the others: It alone finds a contrast in object extraction with vs. without that, and it alone fails to find a contrast in non-inversion with a 3rd-person vs. lexical subject. Otherwise, the three methods reveal exactly the same set of significant contrasts.

Discussion: Our results thus strongly suggest that both categorical and n-point response methods are as reliable as ME in capturing the types of subtle contrasts that are important for syntacticians. Both of these more traditional methods are easy to use, and neither requires significant training or mathematical
sophistication on the part of subjects. If ME has discouraged the more widespread adoption of experimental techniques in syntax, that roadblock may now be removed.

Word count: 749

Y-axis in z-score units. Star between data points = significance (p<0.05).

<table>
<thead>
<tr>
<th>Categorical (yes/no) N=36</th>
<th>5-point (Likert) N=36</th>
<th>Magnitude estimation N=36</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) presence/absence of inversion in wh-questions</td>
<td>= inversion, = no inversion</td>
<td></td>
</tr>
<tr>
<td>(2) that-trace effects = no that, = that</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3) Extraction out of object DP, subject DP and wh-phrase = no extraction, = extraction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4) Effect of subject type on absence of inversion in wh-questions = inversion, = no inversion</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>