Syntactic Judgment Experiments: why to do them and what (not) to worry about

1. Stability of informal judgment collection

   How likely is n-agreement to happen by chance?

   Assuming no bias and independent trials:

   ![Graph showing the results of a sign test with p = 0.05]

   Informal judgment collection is generally effective for robust contrasts. A data set can be assembled quickly.

2. Householder (1973)

   (1) In eliciting judgments, don’t demand pure binary choices, but leave room for a variety of responses. Let all judgments of less than perfect acceptability of clearly understood sentences be specified: does it sound [ … X, Y, Z …]

   (2) Let all sentences be contextualized as sympathetically as possible.

   (3) Use all standard precautions to avoid leading the witness or betraying the problem you’re investigating.

   (4) Rely on your own intuitions only when you are sure they’re universally shared.

   (5) Where possible collect acceptable examples from actual occurrences in texts.

   (6) Make sure your test sentence contains no ‘distractors’.

   (7) For all rejected sentences (if clearly understood), collect preferred equivalents.

   (8) Note any implicational sets that occur.
3. What should you hope to get out of doing an acceptability experiment?
   - Simple orderings of sentence types is least interesting.
   - Gauging the size and interactions of ‘component effects’ is more interesting.
     + e.g., Alexopoulou & Keller on resumption
     Jurka on subject-hood v. freezing
   - Estimating the size and source of (co)variation.
     + e.g., Sprouse on syntactic priming

4. Quantification of systematicity

   **Wh–reciprocals: basic comparisons**

     | Condition       | vs. wh_nm | vs. no_wh | vs. wh_ref |
     |-----------------|-----------|-----------|-----------|
     | wh_ref          |           |           |           |
     | no_wh           |           |           |           |
     | wh_nm           |           |           |           |

   Differences in normalized median ratings by subject
   Positive value indicates preference for wh–reciprocals

   From Bennett (2010) UC Santa Cruz QP.
Fig. 2. Judgments in a that-trace experiment. Each numbered cell represents a different set of four sentences: NT, no that; WT, with that; SE, subject extraction; OE, object extraction. Judgments have been standardized to have a mean of zero and standard deviation of one. The three lines in each cell represent different American universities (adapted from Cowart 1997: 164, figure 36; copyright 1997 by Sage Publications, Inc., reprinted by permission of Sage Publications, Inc., via Copyright Clearance Center).

From Cowart (1997)

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Adj</th>
<th>Arg</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Who can do what about it?</td>
<td>3.73</td>
<td>3.34</td>
<td>4.13</td>
</tr>
<tr>
<td>b. Who can do what about it when?</td>
<td>3.02</td>
<td>2.67</td>
<td>3.36</td>
</tr>
<tr>
<td>c. Who can do what about it, and when?</td>
<td>3.64</td>
<td>3.72</td>
<td>3.57</td>
</tr>
<tr>
<td>d. What can who do about it?</td>
<td>2.27</td>
<td>2.39</td>
<td>2.16</td>
</tr>
<tr>
<td>e. What can who do about it when?</td>
<td>2.27</td>
<td>2.33</td>
<td>2.22</td>
</tr>
<tr>
<td>f. What can who do about it, and when?</td>
<td>2.53</td>
<td>2.52</td>
<td>2.55</td>
</tr>
</tbody>
</table>

6. Your basic menu

   a. Binary/forced-choice ‘acceptable’/’unacceptable’
      a’. speeded
      a”’. pairwise comparison
   b. \(n\)-point/Likert scale
      ordinal
   c. Magnitude Estimation
      intervals

7. Interesting research by Brian Murphy re: the advertised benefits of M.E.
   Language, Interaction and Computation Lab, Centre for Mind/Brain Sciences (CIMeC),
   University of Trento
   ESSLLI slides available here: \url{http://clic.cimec.unitn.it/brian/esslli/}

   Large-scale web surveys
   - ‘authentic’ v. experimentally altered sentences
   - animacy, givenness of indirect object in dative double objects
   - animacy, givenness in passivization

\textit{Calculated rate of convergence of different techniques.}
8. **Atrophy in web experiments**
8. Even the simpler measurement scales can provide rich datasets

9. **Likert scale example**
   Montalbetti sentences (1984): More people have been to Russian than I have
   
   ![Likert scale example](image)

10. **Speeded grammaticality example**    Wagers, Lau & Phillips (in preparation)
   “... the author(s) who the librarian recommend(s) for summer reading ...”
   Kimball & Aissen (1971); Kayne (1983/2006): object agreement in English
   
   ![Speeded grammaticality example](image)
11. Featherston: stop worrying about ‘the finer points’ of data collection

[Diagram showing participant attraction rates with grammatical vs. ungrammatical words and two experiments indicated: Experiment 1 and Experiment 2.]

12. Onto the coarser points ...