Grammaticality Judgments as Linguistic Evidence
Comparing Measurement Scales

Brian Murphy

Language, Interaction and Computation Lab
Centre for Mind/Brain Sciences (CIMeC)
University of Trento

ESSLLI 2009, Bordeaux
Course Outline

- Notions of Grammaticality, and Current Practice in Linguistics
- Others Sources of Linguistic Evidence
- Scales for Measuring Grammaticality
- Guidelines for Grammaticality Judgement Experiments
- Theoretical Implications
Measurement Scales in Social Sciences

- Binary: yes/no, good/bad, black/caucasian
- Categorical: yes/no/don’t know, black/caucasian/latino
- Ranked: good/neutral/bad, Likert scales
- In (experimental) psychology, subjective judgements are largely avoided, but when used, a Likert scale is most common
Magnitude Estimation (ME) I

- Major features of the scale [Stevens, 1975]
  - open ended
  - relative to a baseline sentence (the modulus)
  - continuous, unlimited granularity
  - ratio-based

- Steven’s propounds the Power Law relating the magnitude of stimulus and sensation

- [Bard et al., 1996] propose its use with grammaticality judgements, and use cross-validation method to establish power law
Magnitude Estimation (ME) II

- Origins
  - A (mostly outdated) branch of Psychophysics, the study of behavioural reactions to perception
  - [Stevens, 1946], same guy who came up with nominal/ordinal/interval/ratio data typology in statistics
  - Measuring the precise mathematical relationship between the intensity of physical sensory stimuli (e.g. heat, loudness, brightness) and the perceived sensation
  - Now used principally for quantifying difficult-to-quantify sensations, e.g. taste
Magnitude Estimation (ME) III

- Advertised Advantages
  - adaptive to informant - will be as fine or coarse grained as they and/or the materials demand
  - so, informants not forced to lump sentences together, when they do detect a difference
  - informants do not ‘run out of scale’
  - should eliminate spurious variance
  - increase statistical power
  - ratio scale permits parametric statistics (in contrast to categorial scales)

- ... though ...
  - Unfamiliar, participants find it bizarre, though they get into their stride
  - Common variant demands numerical skills, increased reflection
Magnitude Estimation (ME) IV

Predictions

- Gathering data with magnitude estimation should
  - be an easy task for informants
  - converge more quickly towards a “true” evaluation of sentence grammaticality
  - provide stronger statistical support for previously confirmed theory
Comparing Scales: Case Study I

Participants

▶ Previous ‘benchmark ‘experiment: 48 native English speakers (mostly British Isles)
▶ Comparison of scales: 150 native English speakers (mostly North American)

Materials

▶ 148 dialogue excerpts from popular cinema, including film/scene descriptions
▶ Approx 2/3 authentic, 1/3 experimentally altered; 1/3 test sentences, 2/3 fillers
▶ Investigating productivity of passives and datives
Comparing Scales: Case Study II

Method

- Recruitment and experiment on web
- Participants complete pre-survey
- Acceptability in terms of naturalness
- Blind-assigned to one of three scales:
  - Magnitude Estimation: numerical estimation of ratio acceptability relative to reference sentence: “I can’t go into a film that’s already started” (Annie Hall)
  - Likert Scale: 7 = perfect; 5 = unwieldy; 3-4 ill-formed; 1-2 uninterpretable
  - Pairwise Comparison: is the present sentence better, worse or equally well-formed to the previous item
- First five sentences are practise items
Comparing Scales: Case Study III

- Linguistic phenomenon: the effect of semantic and pragmatic factors on argument structure realisation in English, German, and Chinese

**Actives with Animate Objects**

... I'm warning you

... 

Mean score = 0.15  
(s.d. 0.63)

**Passives with Animate Objects**

... You're being warned by me

... 

Mean score = -0.27  
(s.d. 0.51)

**Actives with Inanimate Objects**

... you hit the wall

... 

Mean score = -0.03  
(s.d. 0.37)

**Passives with Inanimate Objects**

... the wall is hit by you

... 

Mean score = -0.98  
(s.d. 1.34)
Comparing Scales: Case Study IV

- Summary of findings from ‘benchmark’ study (using ME)

<table>
<thead>
<tr>
<th></th>
<th>Chinese 把</th>
<th>Chinese 被</th>
<th>English Passive</th>
<th>German Passive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telicity</td>
<td>yes (+strong)</td>
<td>yes (+strong)</td>
<td>-</td>
<td>yes (+weak)</td>
</tr>
<tr>
<td>Given Object</td>
<td>yes (+weak)</td>
<td>-</td>
<td>-</td>
<td>yes (+weak)</td>
</tr>
<tr>
<td>New Subject</td>
<td>-</td>
<td>yes (+strong)</td>
<td>yes (+weak)</td>
<td>yes (+weak)</td>
</tr>
<tr>
<td>Animate Object</td>
<td>-</td>
<td>-</td>
<td>yes (+weak)</td>
<td>yes (+weak)</td>
</tr>
<tr>
<td>Aware Object</td>
<td>-</td>
<td>yes (-weak)</td>
<td>-</td>
<td>yes (+weak)</td>
</tr>
<tr>
<td>Concrete Object</td>
<td>yes (+weak)</td>
<td>yes (+weak)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Concrete Predicate</td>
<td>yes (+weak)</td>
<td>yes (+strong)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Affected Object</td>
<td>yes (+strong)</td>
<td>yes (+strong)</td>
<td>yes (+weak)</td>
<td>yes (+strong)</td>
</tr>
<tr>
<td>Jackendoff Hierarchy</td>
<td>yes (+weak)</td>
<td>yes (+weak)</td>
<td>-</td>
<td>yes (+weak)</td>
</tr>
</tbody>
</table>

Key: **bold face** $p \leq 0.1$; *italic face* $p > 0.1$
Comparing Scales: Case Study V

Analysis

- Is there a difference in response rates and response times between scales?
- Which scale correlates most closely to a ‘benchmark’ set of judgements?
- Which scale confirms most strongly that:
  - Authentic sentences are more acceptable than experimentally altered ones
  - Animate indirect objects enter into the dative double object construction more readily
    I sent Julien a letter > I sent Bordeaux a letter
  - Given indirect objects enter into the dative double object construction readily
    I sent Julien a letter > I sent a man a letter
  - Animate direct objects enter into the passive construction more readily
  - Context-new subjects enter into the passive more readily
    My wallet was stolen by some &%@! > My wallet was stolen by your &%@! brother
Magnitude Estimation scares away participants!
Comparing Scales: Results II

- ME stabilises faster than pairwise comparisons, slower than Likert scale
- No clear advantage in particular grammaticality ranges

Pearson (linear) correlation over acceptability stimuli between data gathered using three rating scales. Trimmed magnitude estimation data set contains only responses within range of plus/minus three standard deviations.
Comparing Scales: Results III

- ME does not reliably give stronger inferences

<table>
<thead>
<tr>
<th></th>
<th>Benchmark ME ((n=1000))</th>
<th>Comparison ME ((n=1000))</th>
<th>Likert ((n=1000))</th>
<th>Pairwise ((n=1000))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original (55%) vs Altered (45%)</td>
<td>12.5*** ((\infty***))</td>
<td>15.1*** ((\infty***))</td>
<td>18.4*** ((\infty***))</td>
<td>17.4*** ((\infty***))</td>
</tr>
<tr>
<td>Dative animate (50%) vs inanimate (50%) indirect object</td>
<td>4.7*** (3.6****)</td>
<td>2.0* (1.8*)</td>
<td>4.8*** (3.9****)</td>
<td>4.8*** (4.7****)</td>
</tr>
<tr>
<td>Passive animate (50%) vs inanimate (50%) subject</td>
<td>4.0*** (4.5****)</td>
<td>3.8*** (4.2****)</td>
<td>1.9* (1.6)</td>
<td>3.0** (3.1**)</td>
</tr>
<tr>
<td>Dative new (58%) vs given (42%) indirect object</td>
<td>3.3*** (1.7*)</td>
<td>1.1 (1.2)</td>
<td>4.1*** (3.1****)</td>
<td>4.0*** (3.8****)</td>
</tr>
<tr>
<td>Passive given (73%) vs new (27%) subject</td>
<td>2.9*** (2.3*)</td>
<td>4.1*** (3.3****)</td>
<td>1.0 (0.6)</td>
<td>1.8* (1.8*)</td>
</tr>
</tbody>
</table>

Main figures are z-scores from a Welch corrected independent samples t-test over participant responses.

Figures in brackets are z-scores from a Mann–Whitney/Wilcoxon test of independent samples.
Comparing Scales: Results IV

- Corroboration
  - [Fanselow, 2008] find that ME, binary and Likert scales give similar data
  - [Bader and Häussler, ] find that ME, and *speeded* binary judgements give similar data
Comparing Scales: Why? I

- ME assumes a lot: that people can sense differences between sentences that are continuous, open-ended, not absolute; that these differences are ratios; and that people can turn these into numbers.
- People cannot judge grammatical ratios reliably, only intervals.
- Single item not reliable as baseline.
- Logs unmotivated (Featherston, Sprouse).
- Integer preference near zero.
- Numeracy limitations.
- Many practitioners today ignore Steven’s data scale typology.
- Magnitude estimation never worked quite as advertised, even with physical stimuli?
Comparing Scales: Why? II

- Stevens wasn’t a people person?

"How do you expect us to make progress if you make judgments like that!"
A lot of critics *within* psychophysics
Summary I

- Magnitude estimation is generally not the best scale to use for measuring grammaticality
  - At best it performs similarly to other more familiar scales
  - At worst it scares away informants and introduces spurious variance
  - May provide better insight into individual differences (idiolects), but this has not been demonstrated
  - If used, must be closely supervised, with numeracy screening and careful training

- For a general population, a Likert scale is familiar and easy to use, and can be made continuous

- For linguists, a binary or three-way scale may be preferred

- If you prefer to use a binary scale with non-linguists, consider speeded judgements

- Discussion ...
What Comes Next

- Notions of Grammaticality, and Current Practice in Linguistics
- Others Sources of Linguistic Evidence
- Scales for Measuring Grammaticality
- Guidelines for Grammaticality Judgement Experiments
- Theoretical Implications
Alternative Scales

- Featherston: Thermometer scale (in use) Two reference points, at 25% and 75% of “normal scale”
  - Open ended
  - Interval, not ratio
  - Non-continuous, but fine-grained (integers: 20 for lower reference, 30 for upper)

- Five-group grammaticality
  - Arbitrary five-way partition of perfect to ungrammatical (but interpretable) sentences
  - Linguists seem to be able to do this quite reliably

- Another idea [Gobbledygook ..... Perfect] Scale?
References I

Bader, M. and Häussler, J.
Toward a model of grammaticality judgments.
under review.

Magnitude estimation of linguistic acceptability.

Different measures of linguistic acceptability - not so different afterall?

Stevens, S. S. (1946).
On the theory of scales of measurement.

*Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects.*