Five statistical models for Likert-type experimental data on acceptability judgments

Anna Endresen & Laura A. Janda

20.07.2015

Cognitive Linguistics: Empirical Approaches to Russian
A joint project

Anna Endresen
UiT

Laura A. Janda
UiT
Structure of the talk

• Intro
  – What is experimental Likert-type scale data?
  – What statistical tests are appropriate: Parametric vs. Non-parametric

• Our study: marginal change-of-state verbs in Russian
  – Experimental design and our research questions
  – 5 statistical models for collected data
    o ANOVA
    o Ordinal logistic regression
    o Regression mixed-effects model
    o Regression tree & Random forests
    o Classification tree & Random forests

• Conclusions
  – Are the outcomes of these models comparable?
  – Which model is the most appropriate, informative, user-friendly?
Levels of agreement / disagreement

1. Wikipedia has a user friendly interface.
   - strongly agree
   - agree
   - neutral
   - disagree
   - strongly disagree

2. Wikipedia is usually my first resource for research.
   - strongly agree
   - agree
   - neutral
   - disagree
   - strongly disagree

3. Wikipedia pages generally have good images.
   - strongly agree
   - agree
   - neutral
   - disagree
   - strongly disagree

4. Wikipedia allows users to upload pictures easily.
   - strongly agree
   - agree
   - neutral
   - disagree
   - strongly disagree

Rensis Likert (1903-1981)
A method of ascribing quantitative values to qualitative data in order to make it amenable to statistical analysis.

**Likert-type / Likert-like scales** (cf. Lavrakas 2008: 429)

- **Importance**
- **Frequency**
- **Quality**
- **Satisfaction**

### How important are surveys to your startup's success?

- Not at all Important
- Slightly Important
- Moderately Important
- Very Important
- Extremely Important

### How often do you feel in control of your life?
- (1) Never
- (2) Seldom
- (3) Often
- (4) Almost always

### How satisfied are you with our services?

- Very Unsatisfied
- Unsatisfied
- Neutral
- Satisfied
- Very Satisfied
Data measurement: 4 types

• **Nominal scales**
  – Categorize: yes/no, genders, colors, races
  – Convey no quantitative information and no ordering of items

• **Ordinal scales**
  – Order or rank things: movie ranking

• **Interval scales**
  – Order + equal intervals: centimeters, degrees Celsius

• **Ratio scales**
  – Interpretable and natural zero: money, weight, age
Data measurement: 4 types

• **Nominal scales**
  – Categorize: yes/no, genders, colors, races
  – Convey no quantitative information and no ordering of items

• **Ordinal scales**
  – Order or rank things: movie ranking

• **Interval scales**
  – Order + equal intervals: centimeters, degrees Celsius

• **Ratio scales**
  – Interpretable and natural zero: money, weight, age

For nominal and ordinal scales only **non-parametric statistical tests** are appropriate (e.g. chi-squared test).

Interval and ratio scales **allow arithmetic calculations** that compare their sizes, averages, and variation.

**Parametric statistical tests** are appropriate (e.g. ANOVA)
“It is common to treat Likert scales as **interval** level data, it is more conservative to view such data as **ordinal**.” (Lavrakas 2008)

- **Ordinal scales**
  - Order or rank things

- **Interval scales**
  - Order + equal intervals: centimeters, degrees Celsius

- **Ratio scales**
  - Interpretable and natural zero: money, weight, age

Interval and ratio scales **allow arithmetic calculations** that compare their sizes, averages, and variation.

**Parametric statistical tests** are appropriate (e.g. ANOVA)

[Non-parametric statistical tests are appropriate (e.g. chi-squared test).]
The response format (each item is a descriptive statement) is at the **ordinal (categorical)** level of measurement.
The response format (each item is a descriptive statement) is at the **ordinal (categorical)** level of measurement.

The key assumption is that the distances between every two adjacent points are of equal magnitude and justify the **interval** level of measurement. (reinforced graphically or with a set of numbers 1,2,3,4...)
Objections

• They find it illegitimate to use parametric statistics for data obtained via Likert scales.

Likert scales: how to (ab)use them

Susan Jamieson

Dipping my toe into the water of educational research, I have recently used Likert-type rating scales to measure student views on various educational interventions. Likert scales are commonly used to measure attitude, providing ‘a range of responses to a given question or statement’. Typically, there are 5 categories of response, from (for example) 1 = strongly disagree to 5 = strongly agree, although there are arguments in favour of scales with 7 or with an even number of response categories. Between ‘strongly disagree’ and ‘disagree’ is equivalent to the intensity of feeling between other consecutive categories on the Likert scale. The legitimacy of assuming an interval scale for Likert-type categories is an important issue, because the appropriate descriptive and inferential statistics differ for ordinal and interval variables and if the wrong technique is used, the result increases the chance of concluding the wrong conclusion about the significance (or otherwise) of the findings.

The mean (and standard deviation) are inappropriate for ordinal data

The response categories in Likert scales have a rank order, but the intervals between values cannot be presumed equal.
There is more to it: Knapp (1990: 121)

- The distinction between ordinal and interval scales of data measurement is often a challenge when one has to categorize a specific data set.

- A particular scale can be “ordinal, less than ordinal, or more than ordinal”, and that there are no agreed-upon rules for determining this.

- The ordinal / interval scale-and-statistics controversy is a long-standing and continuing debate in the literature. For the history of conflicting views see Gardner 1975.
Possible solution: a different format

<table>
<thead>
<tr>
<th>Question</th>
<th>Rating</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am satisfied with it</td>
<td></td>
<td><strong>Strongly Disagree</strong></td>
</tr>
<tr>
<td>It is simple to use</td>
<td></td>
<td><strong>Strongly Agree</strong></td>
</tr>
<tr>
<td>It is fun to use</td>
<td></td>
<td><strong>Strongly Agree</strong></td>
</tr>
<tr>
<td>It does everything I would expect it to do</td>
<td></td>
<td><strong>Strongly Agree</strong></td>
</tr>
<tr>
<td>I don’t notice any inconsistencies as I use it</td>
<td></td>
<td><strong>Strongly Agree</strong></td>
</tr>
<tr>
<td>It is very user friendly</td>
<td></td>
<td><strong>Strongly Agree</strong></td>
</tr>
</tbody>
</table>
Likert-type scales in linguistics: Elicitation of acceptability judgments

- In linguistic experiments, Likert-type scales are used as a technique for elicitation of acceptability judgments.
- The ambition is to capture the gradient nature of linguistic intuition.
- The subjects are presented with a ranked set of points (usually 5 or 7) where at least the top and the bottom ends are descriptively categorized (cf. Dąbrowska 2010; Bermel and Knittle 2012):

  1  2  3  4  5  6  7

  Unacceptable  | Perfectly normal
Discussion goes on in linguistic studies

• Dąbrowska (2010: 8) points out that a number of studies (Jaccard & Wan 1996; Labovits 1967, Kim 1975) have argued that “parametric tests are quite robust, so that violations of the intervalness assumption have relatively little impact on the results of the test”.

• Dąbrowska 2010 states that “the use of parametric tests with data obtained using Likert scales has now become standard” (cf. similar observations in Blaikie 2003, Pell 2005).

• Dąbrowska (2010) herself uses a Likert-type scale in elicitation experiments and analyzes the responses with ANOVA and t-tests.

• Similarly, Bermel and Knittle (2012a,b) conduct an experiment using a Likert scale and explore their results with ANOVA statistics.
Our study: marginal change-of-state verbs in Russian

• *Ob’jasnit’* ‘clarify, make X be clear’ < *jasnyj* ‘clear_{ADJ}’
• **Two most productive patterns**: o-…-it’ and u-…-it’

• We are interested in **new coinages** like
  – *omuzyskalit’* ‘musicalize’ (< *muzykal’nyj* ‘musical_{ADJ}’)
  – *ukonkretit’* ‘concretize’ (< *konkretnyj* ‘concrete_{ADJ}’)
  – *ovnešnit’* ‘externalize’ (< *vnešnij* ‘external_{ADJ}’)

![Graph showing frequency of word beginnings and endings]
Our experiment:
A score-assignment test

The task: Evaluate the marked word using one of the statements.

Давно пора как-то оприличить наше общение более мягкими выражениями.
‘It’s high time we made our interaction respectable by using kinder statements.’

- 5 points - Это совершенно нормальное слово русского языка.
  ‘This is an absolutely normal Russian word’
- 4 points - Это слово нормальное, но его мало используют.
  ‘This word is normal, but it is rarely used’
- 3 points - Это слово звучит странно, но, может быть, его кто-то использует.
  ‘This word sounds strange, but someone might use it’
- 2 points - Это слово звучит странно, и его вряд ли кто-то использует.
  ‘This word sounds strange and it is unlikely that anyone uses it’
- 1 point - Этого слова в русском языке нет.
  ‘This word does not exist in the Russian language.’
The scale

• The experiment was designed as a score-assignment test.
• Subjects were presented with sixty sentences and a rating system. Each sentence contained a verb that we wanted them to evaluate.
• We used a numerical scale of 5 points combined with a categorical scale of evaluative statements.
• In doing so we wanted to ensure a uniform interpretation of scale points across all subjects.
• What is crucial here is that this 5 point scale is very culturally entranced in Russia because it is commonly used in Russian school and university grades.
• Every Russian is familiar with the scale of 5 points where 5 points mean the best grade.
Our study: on-line questionnaire, one task per page
Stimuli

**STANDARD WORDS**
Words that are standard and conventionalized, might be stored in memory rather than generated on the fly

**MARGINAL WORDS**
Words that are generated by some speakers and can be understood / accepted by some speakers

**NONCE WORDS**
Words that cannot be generated and do not exist

e.g.: undo

e.g.: unworry

e.g.: unblick
Stimuli:
60 prefixed change-of-state verbs with the meaning ‘make X be Y’

**STANDARD WORDS**

10 o + 10 u (control group 1)

**NONCE WORDS**

10 o + 10 u (control group 2)

**MARGINAL WORDS**

10 o + 10 u (experimental group)

- obogatit’ ‘enrich’
- uprostit’ ‘simplify’
- ovnešnit’ ‘externalize’
- uvkusnit’ ‘make tastier’
- očavit’
- učopit’
60 change-of-state verbs as stimuli

• All verbs used in the experiment are deadjectival. This decision is made in order to reduce the number of valuables.

• All standard and marginal verbs chosen for experiment are morphologically transparent and analyzable and have a clear existing adjectival base.

• All stimuli verbs are given as perfective infinitives.
• Verbs are presented in contexts.
  – For standard and marginal change-of-state verbs we use real contexts from the corpus, often shortened.
  – The contexts of nonce verbs are based on corpus contexts of real verbs with meanings similar to those that are assumed for nonce verbs.
3 research questions

PREDICTOR 1: PREFIX
Does the more productive prefix O- form more acceptable novel marginal verbs than the less productive prefix U-?

PREDICTOR 2: AGE OF SPEAKER
Does the speakers’ leniency regarding marginal verbs correlate with age? Do adults (25-62 year old, N=51) have more conservative judgements than children (14-17 year old, N=70)?

PREDICTOR 3: WORD CLASS
Are MARGINAL verbs of the two rival patterns (O- and U-) perceived more like STANDARD or more like NONCE verbs?
Central tendencies in data distribution

Verbs prefixed in O- overall tend to receive higher acceptability scores compared to U-verbs.

Gender does not make any difference.

Children assign higher acceptability ratings than adults.

Marginal verbs received surprisingly low acceptability scores.
## Overview: 5 statistical models

<table>
<thead>
<tr>
<th>Type of test</th>
<th>Name</th>
<th>Type of data</th>
<th>Significant factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parametric</td>
<td>ANOVA</td>
<td>For interval data</td>
<td>WordType</td>
</tr>
<tr>
<td></td>
<td>Ordinal logistic regression</td>
<td>For ordinal data</td>
<td>WordType &gt;&gt;&gt; AgeGroup &gt; Prefix</td>
</tr>
<tr>
<td></td>
<td>Regression mixed-effects model</td>
<td>For ordinal data</td>
<td>WordType &gt;&gt;&gt; AgeGroup</td>
</tr>
<tr>
<td>Non-parametric</td>
<td>Regression tree &amp; Random forests</td>
<td>For numerical ordinal data</td>
<td>WordType &gt;&gt;&gt; AgeGroup &gt; Prefix</td>
</tr>
<tr>
<td></td>
<td>Classification tree &amp; Random forests</td>
<td>For categorical data</td>
<td>WordType &gt;&gt;&gt; Prefix &gt; AgeGroup</td>
</tr>
</tbody>
</table>

**Questions:**

- **Do parametric and non-parametric tests provide different insights?**
- **Which model is the most informative regarding marginal words?**
Model 1. ANOVA

- The impact of **age** and **prefix** is non-significant.

- The impact of **word type** (standard vs. marginal vs. nonce) is significant:
  - F = 546, df = 2, p-value < 2.2e-16

### Parametric test for interval data

<table>
<thead>
<tr>
<th>Type</th>
<th>MAX</th>
<th>MEAN</th>
<th>MIN</th>
<th>stand dev</th>
<th>variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>605</td>
<td>595</td>
<td>549</td>
<td>15</td>
<td>235</td>
</tr>
<tr>
<td>Marginal</td>
<td>479</td>
<td>286.4</td>
<td>169</td>
<td>67</td>
<td>4446</td>
</tr>
<tr>
<td>Nonce</td>
<td>223</td>
<td>183.4</td>
<td>150</td>
<td>19</td>
<td>360</td>
</tr>
</tbody>
</table>
Model 1. ANOVA

T-test RESULTS for **standard vs. marginal** words:
\[ t = 20, \, df = 21, \, p-value = 3.173 \times 10^{-15}, \, 95\% \, confidence \, interval \, is \, 277 - 340 \]

T-test RESULTS for **marginal vs. nonce** words:
\[ t = 7, \, df = 22, \, p-value = 1.098 \times 10^{-6}, \, 95\% \, confidence \, interval \, is \, 71 - 135 \]

- Marginal verbs are evaluated by speakers more like nonce verbs than standard verbs.
- This suggests that speakers are more sensitive to frequency than to semantic transparency.
2. Ordinal logistic regression

- Three factors are statistically significant predictors of acceptability scores – **WordType** and **AgeGroup** and **Prefix**.
- The impact of Gender was found insignificant: Chi-Square= 0.33, df = 1, p-value = 0.56.
- The impact of **WordType** accounts for most of data, while the other two factors are very minor.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Chi-Square</th>
<th>Degrees of freedom</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AgeGroup</td>
<td>59.28</td>
<td>1</td>
<td>&lt; .0001  ***</td>
</tr>
<tr>
<td>Prefix</td>
<td>5.45</td>
<td>1</td>
<td>0.0195   *</td>
</tr>
<tr>
<td>WordType</td>
<td>3415.95</td>
<td>2</td>
<td>&lt; .0001  ***</td>
</tr>
<tr>
<td>TOTAL</td>
<td>3425.06</td>
<td>4</td>
<td>&lt; .0001  ***</td>
</tr>
</tbody>
</table>
Variation across individual marginal stimuli

Variation across subjects

<table>
<thead>
<tr>
<th>Marginal verb</th>
<th>Gloss</th>
<th>Number of subjects who gave</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5 points</td>
<td>4 points</td>
</tr>
<tr>
<td></td>
<td>(“normal word”)</td>
<td></td>
</tr>
<tr>
<td>usovremenit’</td>
<td>22</td>
<td>26</td>
</tr>
<tr>
<td>‘modernize’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>opriličit’</td>
<td>9</td>
<td>25</td>
</tr>
<tr>
<td>‘make decent’</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3. Regression mixed-effects model for ordinal data

<table>
<thead>
<tr>
<th>Random-effects factor</th>
<th>Name</th>
<th>Variance</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SubjectCode</td>
<td>(Intercept)</td>
<td>1.091</td>
<td>1.045</td>
</tr>
<tr>
<td>Stimulus</td>
<td>(Intercept)</td>
<td>1.043</td>
<td>1.021</td>
</tr>
</tbody>
</table>

| Fixed-effects factor        | Estimate | Std.error | z value | Pr(>|z|)       |
|-----------------------------|----------|-----------|---------|----------------|
| AgeGroup-child              | 0.5803   | 0.2013    | 2.883   | 0.00394        |
| WordType-nonce              | -1.7791  | 0.3292    | -5.405  | 6.48e-08       |
| WordType-standard           | 7.4203   | 0.3712    | 19.991  | < 2e-16        |

- The impact of **Gender** and **Prefix** is found insignificant.
- The most optimal fitted model indicated the significant effects of **WordType** and **AgeGroup**.
- The effect of **WordType** is more significant than that of **AgeGroup**.
4 and 5. Regression and Classification trees & Random Forests (CART)

• Classification and Regression Trees is a new method that is quickly gaining popularity in genetics, medicine (Strobl et al. 2009: 324), social sciences, and linguistics (Tagliamonte & Baayen 2012 and Baayen et al. 2013).

• Classification and Regression Trees is a non-parametric statistical technique which is appropriate for non-interval data.

• More flexible in modeling combinations of predictors than Logistic Regression (Faraway 2006: v).

• Trees do not hold any assumptions about the normal distribution of the response variable (as opposed to the logistic regression model)

• Can cope with any data structure and type.

• Highly recommended for unbalanced datasets.
How exactly does CART work?

• CART is an algorithm-based method.
• The outcome of the CART analysis is a graphically plotted “tree” created via a recursive partitioning of data.
• The Tree represents an algorithm of data partitioning which consists of recursive binary splits, each based on one variable.
• The Tree outlines a decision procedure of predicting the values of the dependent variable.
• As a result, recursive splits subdivide the entire data set into several non-overlapping subsets of data.
4. Regression tree & Random forests

Non-parametric test for ordinal data
5. Classification tree & Random forests

Non-parametric test for **categorical** data

- **WordType**
  - p < 0.001
  - {marginal, nonce}
  - standard

- **Prefix**
  - p < 0.001
  - u

- **AgeGroup**
  - p = 0.024
  - adult
  - child

- Node 4 (n = 1210)
  - ABCDE

- Node 6 (n = 510)
  - ABCDE

- Node 8 (n = 700)
  - ABCDE

- Node 9 (n = 1400)
  - ABCDE

- Node 10 (n = 1020)
  - ABCDE

- Node 11 (n = 1210)
  - ABCDE

- Node 12 (n = 1210)
  - ABCDE

- Node 13 (n = 1210)
  - ABCDE
Classification tree & Random forests model

This model demonstrates that the importance of a factor can belong to different “levels”: what is crucial at the level of a local split (AgeGroup and Prefix) might have very small overall predicting power considering the entire dataset, while other factors (like WordType) can determine the major trend of data distribution, as we saw in the major split of the Trees and the highest bar in the Random Forest plots.

The outcome of Random Forest analyses indicates that AgeGroup and Prefix do have some importance but their effect is very small. This effect is revealed in high level interactions of the factors.
Random Forests

Figure 1: Variable importance scale for ordinal data (A>B>C>D>E).

Figure 2: Variable importance scale for categorical data (A, B, C, D, E).

AgeGroup and Prefix do have some importance but their impact on the overall distribution of data is very small.
## Summing up

<table>
<thead>
<tr>
<th>Type of test</th>
<th>Name</th>
<th>Type of data</th>
<th>Significant factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parametric</td>
<td>ANOVA</td>
<td>For interval data</td>
<td>WordType</td>
</tr>
<tr>
<td></td>
<td>Ordinal logistic regression</td>
<td>For ordinal data</td>
<td>WordType &gt;&gt;&gt; AgeGroup &gt; Prefix</td>
</tr>
<tr>
<td></td>
<td>Regression mixed-effects model</td>
<td>For ordinal data</td>
<td>WordType &gt;&gt;&gt; AgeGroup</td>
</tr>
<tr>
<td>Non-parametric</td>
<td>Regression tree &amp; Random forests</td>
<td>For numerical ordinal data</td>
<td>WordType &gt;&gt;&gt; AgeGroup &gt; Prefix</td>
</tr>
<tr>
<td></td>
<td>Classification tree &amp; Random forests</td>
<td>For categorical data</td>
<td>WordType &gt;&gt;&gt; Prefix &gt; AgeGroup</td>
</tr>
</tbody>
</table>

Do parametric and non-parametric tests provide different insights?

Which model is the most informative regarding marginal words?
Conclusions

Do parametric and non-parametric tests provide different insights?

- Parametric tests provide outcomes comparable with non-parametric models:
  - All models identify WordType as the major predictor.
  - The differences concern the factors AgeGroup and Prefix that have very small impact.

Which model is the most informative regarding marginal words?

- Classification Tree & Random Forest model
  - Safest and most appropriate for this data set
  - Most informative
  - Very detailed
  - User-friendly 😊

Anything else?

- The use of culturally entrenched grading scale is an advantage.
Thank you!
Спасибо!
Tusen takk!

Contact us at anna.endresen@uit.no, laura.janda@uit.no
References (1)


References (2)


• Jamieson, S. 2004. Likert scales: How to (ab)use them. In Medical Education. 38. 1212–1218.
References (3)

Additional slides
Impact of Word type

- Types of stimuli (Word type) constitute three distinct patterns in terms of their ratings and MEDIANS.

- Marginal verbs have the MEDIAN score “2” and in this sense they are much closer to nonce verbs (MEDIAN score “1”) than standard verbs.

- Standard verbs, by contrast, receive the MEDIAN score “5” and form the most homogeneous group in terms of ratings.
Whiskers indicate that the ratings of marginal verbs can be as high as score “5” (= “This is an absolutely normal Russian word”)

• Types of stimuli (Word type) constitute three distinct patterns in terms of their ratings and MEDIANS.

• Marginal verbs have the MEDIAN score “2” and in this sense they are much closer to nonce verbs (MEDIAN score “1”) than standard verbs.

• Standard verbs, by contrast, receive the MEDIAN score “5” and form the most homogeneous group in terms of ratings.

Overall marginal verbs received from subjects surprisingly low acceptability scores: half of marginal stimuli received the lowest scores of 1 and 2.

Whiskers of the nonce words reach only the score “3” (= “This word sounds strange, but someone might use it”).
What do these results mean?

• Each type of word has a different behavior
• Marginal words are semantically transparent, but nonce words are not
• **Marginal words** are rated more like nonce words than like standard words
• Speakers are more sensitive to frequency than to semantic transparency
• **Frequency**, which is related to performance, is a stronger factor than competence (ability to unpack morphological patterns)
• **Memory** may be a stronger factor than use of productive rules