Are Readability Formulas Valid Tools for Assessing Survey Question Difficulty?

Timo Lenzner
GESIS – Leibniz Institute for the Social Sciences, Germany

AAPOR Conference 2013, Boston, MA
Readability

- *Readability* refers to readers’ “ease of understanding, owing to the style of writing“ (Klare 2002)

- *Readability formulas* generate numerical estimates of the readability of a text

- Example: Flesch-Kincaid Grade Level index

\[
FKG \text{ score} = 0.39 \times \left( \frac{\text{no. of words}}{\text{no. of sentences}} \right) + 11.8 \times \left( \frac{\text{no. of syllables}}{\text{no. of words}} \right) - 15.59
\]
Readability formulas in survey research

- Converse 1976
- Converse and Schuman 1984
- Gafke and Leuthold 1979
- Harmon 2001
- Holbrook et al. 2007
- Kimball and Kropf 2005
- Velez and Ashworth 2007
- ...
Criticism among linguists

- Based on only two variables
- Word length/frequency and sentence length are not very good predictors of language difficulty (e.g., Anderson and Davison 1988)
- Basic assumptions are violated when applied to short texts, such as survey questions (e.g., Oakland and Lane 2004)
Criterion 1 – Word length

- Assumption: longer words \(\Rightarrow\) more difficult to comprehend

- BUT...

  (1) Current linguistic evidence: word length effect only for pseudo-words and low-frequency words (e.g., Juphard et al. 2004)

  (2) Derivatives: \textit{un-employment, helpless-ness vs. apt, feint}

  (3) Compounds: \textit{safe-guard, over-weight}
Criterion 2 – Word frequency

- Assumption: unfamiliar words \( \Rightarrow \) more difficult to comprehend

- BUT...

  (1) Vocabulary change: word lists become outdated relatively quickly, different socio-cultural groups can have quite different vocabularies

  (2) Semantic transparency: *satisfactory*, *major*, *ear*, *ring* vs. *satisfied*, *majority*, *earring*
Criterion 3 – Sentence length

- Assumption: longer sentences $\Rightarrow$ more difficult to comprehend

- BUT...

  1. Sentence length by itself is not a source of comprehension difficulty (Davison and Green 1988)
  2. Syntactic structures of sentences play a more important role (e.g. left-embedded syntax vs. right-embedded syntax)
Empirical study

- Do four of the most popular readability formulas correctly identify problematic survey questions?
  
  1. Flesch Reading Ease formula (FRE)
  2. Flesch-Kincaid Grade Level index (FKG)
  3. Gunning Fog index (FOG)
  4. Dale-Chall formula (DC)

- 72 question pairs of problematic and improved questions (taken from existing literature on questionnaire design and the Q-Bank database)
Example

Problematic:
Policies that do not safeguard the environment are bad.

Improved:
Policies that safeguard the environment are good.

(Bassili and Scott 1996)
Results: Overall (N=72 questions)

Success rates in % (correct classifications)

- FRE: 51%
- FKG: 50%
- FOG: 39%
- DC: 38%
Results: Example

Problematic:
Policies that do not safeguard the environment are bad.

FRE: 56.7  |  FKG: 7.6  |  FOG: 3.7*  |  DC: 4.2*

Improved:
Policies that safeguard the environment are good.

FRE: 42.6  |  FKG: 9.1  |  FOG: 2.9*  |  DC: 4.1*

(Bassili and Scott 1996)
Results: Agreement between formulas

<table>
<thead>
<tr>
<th>Two-way combinations of formulas</th>
<th>Agreement percent</th>
<th>Kappa (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRE + FKG</td>
<td>90.2</td>
<td>.81 (.07)</td>
</tr>
<tr>
<td>FRE + FOG</td>
<td>56.9</td>
<td>.14 (.11)</td>
</tr>
<tr>
<td>FRE + DC</td>
<td>58.3</td>
<td>.17 (.11)</td>
</tr>
<tr>
<td>FKG + FOG</td>
<td>66.7</td>
<td>.33 (.11)</td>
</tr>
<tr>
<td>FKG + DC</td>
<td>68.1</td>
<td>.36 (.11)</td>
</tr>
<tr>
<td>FOG + DC</td>
<td>95.8</td>
<td>.91 (.05)</td>
</tr>
</tbody>
</table>
Conclusions

- Readability formulas are of limited use in diagnosing survey question difficulty and revising problematic questions
- Formal properties of a text vs. semantic, pragmatic, psycho- and sociolinguistic aspects of language
- Previous studies: correlation vs. causation!
- Limitation: Do the improved question versions indeed produce more reliable and valid data?
  \[\Rightarrow\] Additional research needed
Thanks for your attention!

Contact: timo.lenzner@gesis.org
References

References


Results: Questions from literature (n=15)

Success rates in % (correct classifications)

![Bar chart showing success rates for different categories]

18 May 2013
Lenzner, AAPOR Conference 2013, Boston, MA
Results: Questions from Q-Bank (n=57)

Success rates in % (correct classifications)

![Bar chart showing success rates in % for FRE, FKG, FOG, and DC categories. FRE has the highest success rate at 58%, followed by FKG at 56%, FOG at 42%, and DC at 40%.]
Example 2

Problematic:
How many times have you had bronchitis?

Improved:
How many times have you had bronchitis during the past five years?

(Maitland, Beatty, and Choi 2006)
Results: Example 2

Problematic:
How many times have you had bronchitis?

FRE: 91.0  |  FKG: 2.3  |  FOG: 2.9  |  DC: 4.0

Improved:
How many times have you had bronchitis during the past five years?

FRE: 88.9  |  FKG: 3.8  |  FOG: 4.8  |  DC: 4.3

(Maitland, Beatty, and Choi 2006)