How can variable interrelations be used to detect reversed variable encoding

Miloslav Bahna
Institute for Sociology, Slovak Academy of Sciences

ISSP Research Session, Tampere, May 25, 2014

Note: Research supported by the Slovak Research and Development Agency under grant no. APVV-0309-11.
The problem (TOPBOT and happiness)

ISSP 2011 data, happiness explained by sex, age, subj. health, marital status, religious services attendance and TOP BOT. OLS Beta coefficients
Errata at GESIS archive

- **ISSP 2004 - "Citizenship" - No. 3950**
  Errata of Version 1.0 For Austria, the coding of variable "v301 : R: Top Bottom self-placement 10 pt scale" is completely reversed in version 1, date October 2, 2006 of the data set.

- **ISSP 2006 - "Role of Government IV" - ZA No. 4700**
  2013-04-19: Errata of version v1.0.0 In case of Australia the coding of variable TOPBOT is completely reversed in version v1.0.0 of the data set.

- **ISSP 2007 - "Leisure Time and Sports"**
  2011-3-11 For Australia the coding of variable TOPBOT is completely reversed in version v2.0.0 of the dataset.

- **ISSP 2009 "Social Inequality IV" - ZA No. 5400**
  News and Errata 2012-05-02: In case of United States, the scale is reversed in the variables V44 and TOPBOT.

- **ISSP 2011 - "Health and Health Care" - ZA No. 5800**
  News and Errata 2014-03-11: In case of Slovakia, a user alerted us to an error in the variable TOPBOT. The primary investigator has confirmed that the scale was reversed by mistake. (This is not true!!! :-)}
Solution I.

• All reverse scale errors so far are related to TOPBOT

• Easy to confuse – 1 = TOP? Bottom?

• Initial idea:
  
  – A relative solid cross national positive relationship between TOPBOT and DEGREE could be used to test for correct TOPBOT encoding
ISSP 2009, OLS regression, Beta coefficients, the effect of education on dependent variable subjective social class. Model included the following independent variables: education, ISEI and income explaining social class.
### R: Earnings: South Africa in ZAR

<table>
<thead>
<tr>
<th>Q20 Which social class would you say you belong to?</th>
<th>Mean</th>
<th>N</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower class</td>
<td>5557.97</td>
<td>167</td>
<td>8880.920</td>
</tr>
<tr>
<td>Working class</td>
<td>3092.13</td>
<td>1289</td>
<td>5936.826</td>
</tr>
<tr>
<td>Lower middle class</td>
<td>1518.69</td>
<td>380</td>
<td>2809.530</td>
</tr>
<tr>
<td>Middle class</td>
<td>1004.42</td>
<td>716</td>
<td>2256.238</td>
</tr>
<tr>
<td>Upper middle class</td>
<td>648.10</td>
<td>271</td>
<td>1078.435</td>
</tr>
<tr>
<td>Total</td>
<td>2262.04</td>
<td>2822</td>
<td>4987.056</td>
</tr>
</tbody>
</table>
Solution II.

- Define expected relationships in questionnaire design and test for their existence when depositing data
  - Time consuming
  - Not all relationships are known
We (ISSP) are not the only ones

- **ESS1 Version notes**
  ESS1 edition 5.0 (published 17.06.04):  
  DENMARK: F14 (WRKCTR) In the data file, *values 1 and 2 were reversed* compared to the source questionnaire. The variable has been re-coded by the archive and is now correct.

  ESS1 edition 2.0 (published 21.11.03):  
  Changes from edition 1.0:  
  IRELAND:D25 (IMTCJOB) The question was asked with the *scale reversed* according to the source questionnaire. The variable has been re-coded by the archive for the answer categories to match the other countries.

- **ESS3 Version notes**
  ESS3 edition 3.3 (published 02.02.2011)  
  Changes from edition 3.2:  
  ROMANIA: E48 STFJB and E49 STFJBOT *scale reversed* according to the source questionnaire, data for Romania have been recoded in the international data file. For further details, please see item 46 in the Country Reports in the Documentation Report.
The solution

- Only possible in a cross-national survey: **Let the data determine which variables should be related.**

- If a relationship is negative and significant in 39 countries and positive in the remaining one - something could be (is?) wrong.

- The solution:
  - A script in “R” – spearman correlations are calculated between all numeric variables in the dataset. Suspicious relationships are indicated.
  - All reported problems in ISSP 2009, 2010 and 2011 are found.

- Room for improvement:
  - e.g. testing for relationships between dichotomous variables with a different method.
  - Some false alarms are signaled. Needs fine tuning and expert opinion.
First findings ...

- ISSP 2008
  **GB-GBN-Great Britain** sample – bad MARITAL variable (Average age of Never married, not married = 58.6; Widowed = 54.1)

- ISSP 2009
  **US** – reversed V45 (TOPBOT of respondent's family)
  **Russia** – bad MARITAL variable (mean age of Widowed = 29.8; Never married, single = 65.5)
  ? suspicious relationship SEX – SPWRKST (men have more often full time employed partners than vice versa)
  ? **Argentina** – those who did not vote in last election are older than those who did vote (the only case)
The end of the BOTTOP question