Non-unique Records in International Survey Projects: The Need for Extending Data Quality Control

Kazimierz M. Slomczynski  
The Ohio State University  
Columbus, OH, USA  
and  
Polish Academy of Sciences  
Warsaw, Poland  

Przemek Powałko  
Polish Academy of Sciences  
Warsaw, Poland  

Tadeusz Krauze  
Hofstra University  
Hempstead, NY, USA  

For a given survey data file we define a non-unique record, NUR, as a sequence of all values in a given case (record), which is identical to that of another case in the same dataset. We analyzed 1,721 national surveys in 22 international projects, covering 142 countries and 2.3 million respondents, and found a total of 5,893 NURs concentrated in 162 national surveys, in 17 projects and 80 countries. We show that the probability of the occurrence of any NUR in an average survey sample is exceedingly small, and although NURs constitute a minor fraction of all records, it is unlikely that they are solely the result of random chance. We describe how NURs are distributed across projects, countries, time, modes of data collection, and sampling methods. We demonstrate that NURs diminish data quality and potentially have undesirable effects on the results of statistical analyses. Identifying NURs allows researchers to examine the consequences of their existence in data files. We argue that such records should be flagged in all publically available data archives. We provide a complete list of NURs for all analyzed national surveys.

Keywords: Survey Data Quality; Duplicate Records; Rare Events; Non-Random Errors in Survey Data

1 Introduction

Comparative social sciences rely, to a great extent, on data from international survey projects, usually covering at least a few countries. Specialists in comparative survey methodology produce a large and increasing number of publications on various aspects of data quality (e.g., Biemer & Lyberg, 2003; Gideon, 2012; Harkness, van de Vijver, & Mohler, 2003; Lyberg et al., 1997; McNabb, 2014), for a review of criteria for assessing the quality of cross-national surveys, with references to fitness for intended use, total survey error, and survey process quality, see Survey Research Center, Institute for Social Research, University of Michigan, 2010). However, one aspect of data quality has been largely neglected: the occurrence of non-unique responses across all questions in a given national survey. Although in some books and papers on survey quality “duplicate cases” are referred to as “errors,” systematic assessment of the prevalence of these errors has just begun (Blasius & Thiessen, 2012, 2015; Koczela, Furlong, McCarthy, & Mushtaq, 2015; Kuriakose & Robbins, 2015).

In this paper we explicitly deal with the phenomenon of non-unique records in international social surveys. We find that such records appear in an unexpectedly large proportion of national surveys that used complex questionnaires administered to heterogeneous populations, and were carried out worldwide over the last 50 years.

We start with a definition of non-unique records, and a description of the collection of surveys used in our analysis. After presenting basic findings about the prevalence of such non-unique records, we propose a probabilistic model of a survey, which shows the probability of obtaining duplicates. Following is a more detailed analysis of non-unique records on the level of survey project and country, by time period, and mode of data collection. After discussing implications
of duplicates for results of substantive analyses, we conclude with recommendations for data quality control.

2 Definitions, data and method of identification of NURs

We define a non-unique record (abbreviated as NUR) as a sequence of all values of variables comprising a given case (record), which is identical to that of another case in the same dataset. In the language of survey methodology, a NUR corresponds to a sequence of all answers (including non-responses) given by a respondent, which is identical to that of another respondent in the same national survey. In the literature, such records are known as duplicates. The concept of duplicates may be misleading because it suggests that there is an original that has been duplicated. However, given two identical records of respondents’ answers, it is not possible to determine which record is the original one – at least not without external information. For this reason we prefer to use the concept of NUR and refer to a “duplicate record” as its synonym.

We apply the concept of NUR to a collection of 1,721 national surveys in 22 projects covering 142 countries or territories, and 2.3 million respondents, during the period 1966-2013. International projects were chosen according to the following criteria: (a) the projects are non-commercial; (b) they were designed as cross-national, and – preferably – multi-wave; (c) the samples were intended to represent territories, and 2.3 million respondents, during the period – preferably another respondent in the same national survey. In the literature, such records are known as duplicates. The concept of duplicates may be misleading because it suggests that there is an original that has been duplicated. However, given two identical records of respondents’ answers, it is not possible to determine which record is the original one – at least not without external information. For this reason we prefer to use the concept of NUR and refer to a “duplicate record” as its synonym.

In order to obtain records of values of variables corresponding to questionnaire items, that is, to questions to which respondents were providing answers during the interview, the following types of variables have been excluded from the original datasets: (a) technical variables (e.g., variables created at the administrative level, e.g., population/post-stratification weights, geographical regions, size/type of community), (b) variables containing interviewers’ remarks (e.g., interview details, level of respondent’s cooperation, respondent’s race), (c) variables derived from respondents’ answers (e.g., BMI, classifications of education/occupational levels), and (d) all variables which can be derived from sample characteristics or from the construction of the sample (e.g., respondents’ age and gender, and information about household members).

The method of finding NURs consisted of pairwise comparisons of each case with every other case within a given national survey dataset. Response options among the considered variables ranged from dichotomous to hundreds of categories, and comparisons were done on raw values of all variables, which include both codes for substantive responses and missing values. We have chosen this procedure because it allowed us to establish distributions of similarities for which NURs are extreme cases (perfect matches). An alternative, and much faster procedure, would be a simple sorting of all records in a dataset and comparing neighboring records; however, it would not provide information on how NURs differ from other similar cases. A study of these similarities is outside the scope of this paper.

3 Basic findings

From among 1,721 national surveys, 162 surveys (9.4% of the total) in 17 projects contain a total of 5,893 NURs (see Table 2). In 52% of the affected surveys a single duplicate record was found. In the remaining 48% we found several patterns of NURs, such as multiple doublets or records repeated three, four, or even more times, often in combination. For example, a survey conducted in Ecuador (Latino-barómetro, 2000), contains the largest number of 733 NURs (i.e., 272 doublets and 63 triplets) in the sample of 1,200, which means that over 60% of records are non-unique. An example of a survey with the most diverse pattern of NURs comes from Norway (ISSP, 2009), and has 54 NURs in 27 doublets, 36 in 12 triplets, 24 in 6 quadruplets, 25 in 5 quintuplets, 6 in 1 sextuplet, 7 in 1 septuplet, and 8 in 1 octuplet, with a total of 160 NURs in the sample of 1,456 (11.0%). Among total, in the 5,893 NURs, 5,232 are doublets, 393 are triplets, 188 are quadruplets, 30 quintuplets, 12 sextuplets, one septuplet and one octuplet, as well as a single record repeated 23 times.

1Links to the used source data files, as well as all documentation allowing for complete replication of the analysis are available in supplementary materials. We provide four types of files: (1) files needed for preparation of source datasets (pub-1-general info.xlsx, pub-2-sources of datatiles-v2.xlsx, pub-3_README.docx, pub-3_correcting and converting files.docx, pub-3-merging and patching EVS and WVS.docx, pub-3-merging EB.docx, pub-3-merging ESS.docx, pub-3-merging ISSP.docx); (2) files needed for identification of duplicates (pub-4-variables taken into account.docx, pub-5-IDs of duplicates.xlsx); (3) Stata data file with variables used in this paper (NURs.dta); (4) statistical procedures for obtaining the results presented in Tables 1 to 8 (pub-6-statistical procedures.docx).

2For the complete list of NURs see pub-5-IDs of duplicates.xlsx in supplementary materials. Among NURs, only 67 are clearly lacking the respondents’ answers as if the relevant interviews had been interrupted or not even begun.
Table 1

<table>
<thead>
<tr>
<th>Survey project</th>
<th>Time span</th>
<th>Number of surveys</th>
<th>Number of distinct countries</th>
<th>Average number of questions</th>
<th>Average sample size</th>
<th>Number of records</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABS</td>
<td>2001-2011</td>
<td>30</td>
<td>13</td>
<td>174</td>
<td>1456</td>
<td>43691</td>
</tr>
<tr>
<td>AFB</td>
<td>1999-2009</td>
<td>66</td>
<td>20</td>
<td>210</td>
<td>1499</td>
<td>98942</td>
</tr>
<tr>
<td>AMB</td>
<td>2004-2012</td>
<td>92</td>
<td>24</td>
<td>178</td>
<td>1645</td>
<td>151341</td>
</tr>
<tr>
<td>ARB</td>
<td>2006-2011</td>
<td>16</td>
<td>11</td>
<td>219</td>
<td>1230</td>
<td>19684</td>
</tr>
<tr>
<td>ASES</td>
<td>2000</td>
<td>18</td>
<td>18</td>
<td>193</td>
<td>1014</td>
<td>18253</td>
</tr>
<tr>
<td>CB</td>
<td>2009-2012</td>
<td>12</td>
<td>3</td>
<td>275</td>
<td>2052</td>
<td>24621</td>
</tr>
<tr>
<td>CDCEE</td>
<td>1990-2001</td>
<td>27</td>
<td>16</td>
<td>299</td>
<td>1071</td>
<td>28926</td>
</tr>
<tr>
<td>CNEP&lt;sup&gt;d&lt;/sup&gt;</td>
<td>2004-2006</td>
<td>8</td>
<td>8</td>
<td>294</td>
<td>1672</td>
<td>13372</td>
</tr>
<tr>
<td>EB&lt;sup&gt;c&lt;/sup&gt;</td>
<td>1983-2012</td>
<td>152</td>
<td>37</td>
<td>342</td>
<td>913</td>
<td>138753</td>
</tr>
<tr>
<td>EQLS</td>
<td>2003-2012</td>
<td>93</td>
<td>35</td>
<td>167</td>
<td>1315</td>
<td>105527</td>
</tr>
<tr>
<td>ESS</td>
<td>2002-2013</td>
<td>146</td>
<td>32</td>
<td>223</td>
<td>1928</td>
<td>281496</td>
</tr>
<tr>
<td>EVS</td>
<td>1981-2009</td>
<td>128</td>
<td>50</td>
<td>347</td>
<td>1301</td>
<td>166502</td>
</tr>
<tr>
<td>ISJP</td>
<td>1991-1997</td>
<td>21</td>
<td>14</td>
<td>205</td>
<td>1229</td>
<td>25805</td>
</tr>
<tr>
<td>ISSP&lt;sup&gt;c&lt;/sup&gt;</td>
<td>1985-2013</td>
<td>363</td>
<td>53</td>
<td>88</td>
<td>1359</td>
<td>493243</td>
</tr>
<tr>
<td>LB</td>
<td>1995-2010</td>
<td>260</td>
<td>19</td>
<td>251</td>
<td>1134</td>
<td>294965</td>
</tr>
<tr>
<td>LITS</td>
<td>2006-2010</td>
<td>64</td>
<td>35</td>
<td>636</td>
<td>1060</td>
<td>67866</td>
</tr>
<tr>
<td>NBB</td>
<td>1993-2004</td>
<td>18</td>
<td>3</td>
<td>172</td>
<td>1200</td>
<td>21601</td>
</tr>
<tr>
<td>PA2&lt;sup&gt;d&lt;/sup&gt;</td>
<td>1979-1981</td>
<td>3</td>
<td>3</td>
<td>271</td>
<td>1352</td>
<td>4057</td>
</tr>
<tr>
<td>PA8NS</td>
<td>1973-1976</td>
<td>8</td>
<td>8</td>
<td>345</td>
<td>1574</td>
<td>12588</td>
</tr>
<tr>
<td>PPE7N</td>
<td>1966-1971</td>
<td>7</td>
<td>7</td>
<td>299</td>
<td>2360</td>
<td>16522</td>
</tr>
<tr>
<td>VPCPCE&lt;sup&gt;d&lt;/sup&gt;</td>
<td>1993</td>
<td>5</td>
<td>5</td>
<td>193</td>
<td>945</td>
<td>4723</td>
</tr>
<tr>
<td>WVS</td>
<td>1981-2008</td>
<td>184</td>
<td>89</td>
<td>221</td>
<td>1394</td>
<td>256582</td>
</tr>
<tr>
<td>Total</td>
<td>1966-2013</td>
<td>1721</td>
<td>142</td>
<td>228</td>
<td>1330</td>
<td>2289060</td>
</tr>
</tbody>
</table>

<sup>a</sup>Data were downloaded at the turn of 2013/2014. For detailed dates and links to data sources, see supplementary materials.

<sup>b</sup>Countries or territories.

<sup>c</sup>For CNEP, EB, and ISSP, only selected survey editions were used.

<sup>d</sup>For CNEP, PA2, and VPCPCE, numbers come from the source files after filtering out panel and post-election surveys.

Abbreviations: Asian Barometer (ABS), Afrobarometer (AFB), Americas Barometer (AMB), Arab Barometer (ARB), Comparative National Elections Project (CNEP), Asia Europe Survey (ASES), Caucasus Barometer (CB), Consolidation of Democracy in Central and Eastern Europe (CDCEE), Eurobarometer (EB), European Quality of Life Survey (EQLS), European Social Survey (ESS), European Values Study (EVS), International Social Justice Project (ISJP), International Social Survey Programme (ISSP), Latinobarometro (LB), Life in Transition Survey (LITS), New Baltic Barometer (NBB), Political Action II (PA2), Political Action - An Eight Nation Study (PA8NS), Values and Political Change in Postcommunist Europe (VPCPCE), Political Participation and Equality in Seven Nations (PPE7N), World Values Survey (WVS).

4 Probabilistic model

In order to evaluate the probability of the occurrence of NURs we formulate a mathematical model. The probability of a single duplicate, that is two NURs, is equal to the probability of two respondents in the same survey providing the same answers to all questions. This probability is determined by the number of respondents, the number of questions, the number of response categories, and the dependence among answers to different questions.

Average sample sizes in the projects from our collection range from 913 to 2,360, with a global average equal to 1,330 (see Table 1). The average number of questions addressed to a respondent in the survey questionnaires ranges from 88 to 636, with the global average of 228. To estimate the probability of duplicate records, we assume dichotomous variables (binary choices) with equal probabilities of both values, and the statistical independence of answers to one-third of the
Table 2
17 International Survey Projects with Non-unique Records, Ordered by the Percent of Countries with NURs

<table>
<thead>
<tr>
<th>Survey project</th>
<th>Number of surveys with NURs</th>
<th>Number of surveys</th>
<th>Number of surveys in NURs</th>
<th>Percent of surveys</th>
<th>Percent of surveys with NURs</th>
<th>Number of surveys with NURs</th>
<th>Percent of surveys with NURs</th>
</tr>
</thead>
<tbody>
<tr>
<td>LB</td>
<td>13</td>
<td>68.42</td>
<td>12.31</td>
<td>1225</td>
<td>0.42</td>
<td>35633</td>
<td>3.44</td>
</tr>
<tr>
<td>AMB</td>
<td>10</td>
<td>41.67</td>
<td>32</td>
<td>12.94</td>
<td>0.03</td>
<td>22431</td>
<td>0.21</td>
</tr>
<tr>
<td>ISSP</td>
<td>19</td>
<td>35.85</td>
<td>31</td>
<td>8.54</td>
<td>0.19</td>
<td>59587</td>
<td>1.55</td>
</tr>
<tr>
<td>WVS</td>
<td>31</td>
<td>34.83</td>
<td>36</td>
<td>19.57</td>
<td>0.77</td>
<td>54449</td>
<td>3.62</td>
</tr>
<tr>
<td>CB</td>
<td>1</td>
<td>33.33</td>
<td>1</td>
<td>8.33</td>
<td>0.01</td>
<td>1975</td>
<td>0.10</td>
</tr>
<tr>
<td>NBB</td>
<td>1</td>
<td>33.33</td>
<td>1</td>
<td>5.56</td>
<td>0.01</td>
<td>1987</td>
<td>0.10</td>
</tr>
<tr>
<td>ABS</td>
<td>3</td>
<td>23.08</td>
<td>3</td>
<td>10.00</td>
<td>0.03</td>
<td>7289</td>
<td>0.16</td>
</tr>
<tr>
<td>EB</td>
<td>8</td>
<td>21.62</td>
<td>11</td>
<td>7.24</td>
<td>0.57</td>
<td>10773</td>
<td>7.40</td>
</tr>
<tr>
<td>LITS</td>
<td>7</td>
<td>20.00</td>
<td>7</td>
<td>10.94</td>
<td>0.05</td>
<td>7001</td>
<td>0.46</td>
</tr>
<tr>
<td>EQLS</td>
<td>7</td>
<td>20.00</td>
<td>8</td>
<td>8.60</td>
<td>0.04</td>
<td>8549</td>
<td>0.47</td>
</tr>
<tr>
<td>AFB</td>
<td>4</td>
<td>20.00</td>
<td>4</td>
<td>6.06</td>
<td>0.03</td>
<td>9092</td>
<td>0.31</td>
</tr>
<tr>
<td>CDCEE</td>
<td>3</td>
<td>18.75</td>
<td>3</td>
<td>11.11</td>
<td>0.58</td>
<td>3740</td>
<td>4.49</td>
</tr>
<tr>
<td>ESS</td>
<td>5</td>
<td>15.63</td>
<td>5</td>
<td>3.42</td>
<td>0.00</td>
<td>10227</td>
<td>0.14</td>
</tr>
<tr>
<td>PPE7N</td>
<td>1</td>
<td>14.29</td>
<td>1</td>
<td>14.29</td>
<td>0.31</td>
<td>1769</td>
<td>2.94</td>
</tr>
<tr>
<td>EVS</td>
<td>5</td>
<td>10.00</td>
<td>5</td>
<td>3.91</td>
<td>0.34</td>
<td>10224</td>
<td>5.58</td>
</tr>
<tr>
<td>ISJP</td>
<td>1</td>
<td>7.14</td>
<td>1</td>
<td>4.76</td>
<td>0.01</td>
<td>1001</td>
<td>0.20</td>
</tr>
<tr>
<td>ASES</td>
<td>1</td>
<td>5.56</td>
<td>1</td>
<td>5.56</td>
<td>0.04</td>
<td>1000</td>
<td>0.80</td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
<td>56.34</td>
<td>162</td>
<td>9.41</td>
<td>0.26</td>
<td>246727</td>
<td>2.39</td>
</tr>
</tbody>
</table>

The uniqueness of records under the above assumptions is considered in terms of the classical birthday problem concerning the probability that among a given number of persons there will be a pair with the same birthday (Feller, 1968, p. 33). In our case, the birthday problem is modified by replacing the number of days in a year by the number of possible sets of answers.

The probability \( p \) of obtaining at least one duplicate within \( k \) independent binary variables, given a sample size of \( N_r \), where \( N_r \ll 2^k \), can be approximated by

\[
p \approx 1 - \exp\left(-\frac{N_r^2}{2^{k+1}}\right)
\]

Probabilities for realistic numbers of respondents and independent variables are presented in Figure 1. For example, for the average sample size of surveys in our collection, \( N_r = 1,330 \), and the number of independent variables \( k \) ranging from 30 to 60, the probability \( p \) varies from \( 8.23 \cdot 10^{-4} \) to a low of \( 7.67 \cdot 10^{-13} \), which demonstrates the unlikeliness of duplicates under assumptions of this simple model.

The number of respondents \( N_r \) required for obtaining a single duplicate, resulting from the reformulation of the above equation, is

\[
N_r \approx \sqrt{-2^{k+1} \log (1 - p)}
\]

![Figure 1. Estimated Probability of at least One Duplicate by Sample Size and Number of Independent Binary Variables (k)](attachment:image.png)
Applying this equation to the data in Table 1 shows that, for example, for 76 independent binary variables (one-third of the average number of questions per national survey, i.e., 228) one would need $3.90 \cdot 10^{10}$ respondents in order to find a pair of identical sets of answers with the probability 0.01. In the case of one-third of the lower (88) and upper (636) bounds of the numbers of questionnaire items (i.e., respectively 29 and 212), the numbers of respondents needed for a duplicate are, respectively, 3,285 and 1.15 $\cdot 10^{31}$ (with the same probability 0.01). Even though the number 3,285 sounds realistic as a sample size, we should remember that we would still need 100 samples of this size to expect a single duplicate in one of them, for as few as 29 independent variables. The intuitive understanding of the model can be based on the fact that the order of magnitude of the number of respondents ($N_r \approx 10^3$) is much smaller than the order of magnitude of possible response patterns ($N_p \approx 10^{22}$ for one-third of the average number of questions per national survey, i.e., 76 independent variables). As a result, one would not expect to encounter any NURs in surveys carried out under these model assumptions.

How would the violation of the assumptions of binary variables and the independence of one-third of variables affect estimates obtained from the above model? The assumption about dichotomous answers provides the basis for a conservative estimate, since in practice respondents’ answers are coded in multiple categories, which makes a duplicate record much less probable. The assumption of independence for a subset of questionnaire items is supported empirically: the usual pattern of statistically significant correlations of respondents’ answers for a typical survey suggests that violations of postulated independence for one-third of items occur only rarely.\(^3\) One should take into account that a lower share of independent variables increases the probability of obtaining a duplicate, while a larger number of response categories has the opposite effect. In this context, we note that under our model even with a small number of independent items, if these items are multi-category responses and as such expressible as sets of binary variables, the probability of obtaining NURs would be comparable to those calculated above. For this reason we claim that our simple model is adequate for analyzing NURs in the international survey projects listed in Table 1. However, a more universal model, also applicable to special populations and one-theme-focused questionnaires, should take into account additional conditions (for discussion of these issues see Simmons, Mercer, Schwarzer, & Kennedy, 2016).

5 Correlates of NURs

The above probabilistic model describes the likelihood of obtaining NURs by chance, and shows that such occurrences are very unlikely. However, as we had shown earlier, not only single duplicates, but complex patterns of NURs, are common and universal. In the following section we analyze the incidence of NURs in various aspects, in an attempt to identify patterns of variation that could bring us closer to understanding the mechanisms that create NURs in international social surveys.

As shown in Figure 2, the degree of inequality in the number of NURs among the 162 affected surveys, is considerable: 80% of all NURs (i.e. 4,735 out of the total 5,893) are present in just 14 surveys, while the remaining 148 surveys contain 20% of NURs.

This differentiation motivates further investigation. A particular survey is identified by the project, country and year. We address these three aspects in that order, followed by a discussion of the variation across survey modes and sampling methods.

5.1 Survey projects

The distribution of NURs across the 17 affected survey projects is not uniform (see Table 2). For example, NURs appear in 19.6% of surveys of the World Values Survey (the highest share) and 3.4% of surveys of the European Social Survey (the lowest share). Additionally, within each project, there are differences with respect to the number of countries in which surveys have NURs. In the extreme case, surveys in 13 out of 19 countries included in Latinobarometro contain NURs.

\(^3\)This empirical evidence gives only plausible support for our assumption since even zero-correlations do not imply statistical (stochastic) independence.
Six projects contain surveys with at least 10% of NURs: Consolidation of Democracy in Central and Eastern Europe, Eurobarometer, European Values Study, International Social Survey Programme, Latinobarometro, and World Values Survey. In the most extreme case of a survey in Latinobarometro in Ecuador in the year 2000, over 60% of the sample consists of NURs. For all 14 surveys see Table 3. We examined these surveys to be sure that NURs are not produced by an excess of missing data or by the specific structure of questionnaires.

5.2 Countries

Overall, national surveys in 80 out of 142 countries have NURs. These countries differ considerably in terms of the number of surveys. In our collection we have 38 countries and territories with one, two, or three surveys, resulting in a total of 76 surveys. In eight of these countries we have only one survey with NURs. In the remaining 77 countries, the number of surveys with NURs ranges from 1 out of 30 (Sweden) to 6 out of 25 (Portugal). In most of these countries, the number of NURs per survey is relatively small, ranging from 0.01 to 1.51. However, in five countries (United States, Mexico, Belgium, Norway, and Austria), the duplication rate exceeds 100 NURs per affected survey. A comparison of the maximal number of NURs with the number of NURs per survey with NURs indicates the concentration of NURs. For example, in the United States and Norway all NURs occur in a single survey, while in Bulgaria NURs are spread over six surveys, although only one of them is particularly troublesome in terms of concentration of NURs (as shown in the last column in Table 4). We observe that NURs were found in countries at all levels of economic development (e.g., Japan, Mexico, and Ethiopia) and with different political systems (e.g., Norway, Romania, and Panama).

5.3 Time

The rapid growth of NURs begins in 1981 (see Figure 3 and Table 5). Till that time we found only one survey with NURs among 17 surveys analyzed. In the period of 1981-1996 we found 30 surveys with NURs, 10.56% of the total; in terms of records this corresponds to 0.84% for all records and 6.60% for records in surveys with NURs. In 1996, in the cumulative distribution, 50% of NURs corresponds to 17% of all records (or surveys).

On the basis of Figure 3 we may distinguish two other specific periods: 1997-2005 and 2006-2013. In the first of these periods the proportion of affected surveys is still above 10%, but in terms of records the increase of NURs slows down. We note that proportion of NURs is 0.24% among all records and 1.81% among records in the affected surveys. At the end of this period 80% of all NURs appears in 50% of all records (surveys). In the last period 2006-2013 the process slows down even more: remaining 20% of NURs corresponds to 50% of all records (surveys). At this period proportion of NURs is 0.10% among all records and 1.11% among records in the affected surveys.

The insert in Figure 3 illustrates the role of 14 surveys with the highest proportion of NURs. Even if these surveys are excluded, the tendency of NURs growth is the same: the cumulative proportion of NURs is larger than the cumulative proportion of all records (surveys) and the two curves meet only at a single time point (year 1996).

5.4 Mode of data collection

It is reasonable to expect that the occurrence of NURs is related to a specific mode of data collection. In Table 6 we provide data limited to three survey projects that include the largest number of NURs and frequently document the modes of data collection (International Social Survey Programme, Latinobarometro, World Values Survey). The majority of surveys used face-to-face interviews, of which most failed to specify the exact mode of data recording. Of the 444 surveys in the “face-to-face, not specified” category, 70 surveys contain 3,702 NURs. Among surveys with a specified PAPI/CAPI mode, the percentages of NURs are around 0.06. Survey modes are not randomly distributed across survey projects, hence the high proportion of NURs in this group might as well be attributed to the survey project as to the mode effect.

Far fewer surveys used self-completion questionnaires, and in this group the share of NURs ranges from 0 to 0.03 percent. The case of mixed mode (mail/web in Norway, ISSP 2009) was selected because of an interesting feature: all NURs in web questionnaires have non-unique counterparts in the mailed-back mode. This is a puzzling example of cross-mode NURs.

5.5 Sampling methods

Following Kohler (2007) we employ his classification of sampling methods (“simple/stratified random sampling”, “multistage individual register”, “multistage address register”, “multistage random route”, “multistage unspecified”, “quota”). All documentation of 22 international survey projects was examined with respect to description of sampling methods using keywords. First, the quota sample was

4Here we consider states, thus the territories of Belgium-Wallonia in ISSP 2011 and Russia-Krasnoyarsk in CDCEE 1, which also have identified NURs, are omitted.
Table 3
14 National Surveys with the Largest Proportion of NURs, Ordered by the Percent of NURs

<table>
<thead>
<tr>
<th>Survey project / wave</th>
<th>Country</th>
<th>Number of records</th>
<th>Number of NURs</th>
<th>Percent of NURs</th>
</tr>
</thead>
<tbody>
<tr>
<td>LB/2000</td>
<td>Ecuador (EC)</td>
<td>1200</td>
<td>733</td>
<td>61.08</td>
</tr>
<tr>
<td>WVS/5</td>
<td>Ethiopia (ET)</td>
<td>1500</td>
<td>539</td>
<td>35.93</td>
</tr>
<tr>
<td>EB/21</td>
<td>Belgium (BE)</td>
<td>1018</td>
<td>344</td>
<td>33.79</td>
</tr>
<tr>
<td>LB/1996</td>
<td>Panama (PA)</td>
<td>1005</td>
<td>316</td>
<td>31.44</td>
</tr>
<tr>
<td>WVS/5</td>
<td>South Korea (KR)</td>
<td>1200</td>
<td>354</td>
<td>29.50</td>
</tr>
<tr>
<td>EVS/1</td>
<td>United States (US)</td>
<td>2325</td>
<td>528</td>
<td>22.71</td>
</tr>
<tr>
<td>WVS/3</td>
<td>Mexico (MX)</td>
<td>2364</td>
<td>537</td>
<td>22.72</td>
</tr>
<tr>
<td>EB/31</td>
<td>Belgium (BE)</td>
<td>1002</td>
<td>220</td>
<td>21.96</td>
</tr>
<tr>
<td>ISSP/1989</td>
<td>Austria (AT)</td>
<td>1997</td>
<td>374</td>
<td>18.73</td>
</tr>
<tr>
<td>WVS/1</td>
<td>Japan (JP)</td>
<td>1204</td>
<td>195</td>
<td>16.20</td>
</tr>
<tr>
<td>EB/19</td>
<td>Belgium (BE)</td>
<td>1038</td>
<td>148</td>
<td>14.26</td>
</tr>
<tr>
<td>CDCEE/1</td>
<td>Romania (RO)</td>
<td>1234</td>
<td>154</td>
<td>12.48</td>
</tr>
<tr>
<td>ISSP/1998</td>
<td>Bulgaria (BG)</td>
<td>1102</td>
<td>133</td>
<td>12.07</td>
</tr>
<tr>
<td>ISSP/2009</td>
<td>Norway (NO)</td>
<td>1456</td>
<td>160</td>
<td>10.99</td>
</tr>
</tbody>
</table>

Figure 3. Cumulative Percent Distribution of National Surveys, Records and NURs. Percent of national surveys (blue line) covers 1,721 surveys. The marked years here are those in which the largest increase of NURs occurred. The first number is the number of surveys with NURs in the given year; the second one (in bold) is the total number of surveys in this year. Percent of records (black line) covers all 2,289,060 records; it closely fits the survey’s line. Percent of NURs (red line) covers 5,893 NURs. Dots refer to the points of the largest increases of NURs. The number in bold is the total number of NURs that occurred in all surveys in the corresponding years. For these years, 14 national surveys with the largest proportion of NURs are marked, with number of NURs that they contain. For identification of national surveys see Table 3. Insert figure shows the NURs’ line (in red) and all records line (in black) after removing 14 national surveys with the largest proportion of NURs. The closest distance between the two curves occurs in 1996 at the 17 percent level.
identified by keywords “quota” and its equivalent in Spanish. Next, the search included “route” and “walk” to filter out the method “multistage random route”. Other methods were also determined by appropriate keywords. All assigned methods were verified by reading the entire description of sampling methods for each national survey. We added the category of “insufficient information” (for description lacking details), combining it with “no information” (in case of missing description of sampling method) and Kohler’s “multistage unspecified” (when the type of multistage sampling was not identified).

In terms of surveys with NURs, quota sampling and the cases with insufficient information are the worst; in both these categories the proportion of surveys with NURs exceeds 10%. We should add that in these two categories there are 12 national surveys with the largest number of NURs (out of all 14). These two categories are also the worst in terms of proportion of NURs: for quota this proportion is 0.60%, and for insufficient information – 0.3%. The best are multistage random route sampling and multistage individual register sampling with proportion 0.04% and 0.05%, respectively. Ordering of sampling methods in terms of percentage of NURs among all records is the same as the ordering of sampling methods with respect of density of NURs in affected surveys.

In the last column of Table 7 we show that proportion of records with NURs in affected surveys is particularly high in the case of quota sampling and the case of insufficient information about sampling; however it is not negligible in the case of other sampling methods, in the case of multistage address register reaching almost 3%.

![Table 4](image)

<table>
<thead>
<tr>
<th>Country</th>
<th>Number of surveys</th>
<th>Number of surveys with NURS</th>
<th>Percent of surveys with NURS</th>
<th>Number of records</th>
<th>Number of NURs</th>
<th>Percent of NURs</th>
<th>Average number of NURs per survey with NURs</th>
<th>Maximal number of NURs in a survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portugal</td>
<td>25</td>
<td>6</td>
<td>24.00</td>
<td>35700</td>
<td>74</td>
<td>0.21</td>
<td>12.33</td>
<td>40</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>28</td>
<td>6</td>
<td>21.43</td>
<td>34384</td>
<td>146</td>
<td>0.42</td>
<td>24.33</td>
<td>133</td>
</tr>
<tr>
<td>Belgium</td>
<td>20</td>
<td>4</td>
<td>20.00</td>
<td>28400</td>
<td>714</td>
<td>2.51</td>
<td>178.50</td>
<td>344</td>
</tr>
<tr>
<td>Guatemala</td>
<td>20</td>
<td>3</td>
<td>15.00</td>
<td>22755</td>
<td>24</td>
<td>0.11</td>
<td>8.00</td>
<td>20</td>
</tr>
<tr>
<td>El Salvador</td>
<td>20</td>
<td>3</td>
<td>15.00</td>
<td>23234</td>
<td>6</td>
<td>0.03</td>
<td>2.00</td>
<td>2</td>
</tr>
<tr>
<td>Ireland</td>
<td>30</td>
<td>4</td>
<td>13.33</td>
<td>39555</td>
<td>20</td>
<td>0.05</td>
<td>5.00</td>
<td>8</td>
</tr>
<tr>
<td>Venezuela</td>
<td>23</td>
<td>3</td>
<td>13.04</td>
<td>28185</td>
<td>70</td>
<td>0.25</td>
<td>23.33</td>
<td>60</td>
</tr>
<tr>
<td>Austria</td>
<td>24</td>
<td>3</td>
<td>12.50</td>
<td>31923</td>
<td>430</td>
<td>1.35</td>
<td>143.33</td>
<td>374</td>
</tr>
<tr>
<td>Argentina</td>
<td>24</td>
<td>3</td>
<td>12.50</td>
<td>28769</td>
<td>32</td>
<td>0.11</td>
<td>10.67</td>
<td>28</td>
</tr>
<tr>
<td>Russia</td>
<td>25</td>
<td>3</td>
<td>12.00</td>
<td>46871</td>
<td>10</td>
<td>0.02</td>
<td>3.33</td>
<td>4</td>
</tr>
<tr>
<td>Denmark</td>
<td>27</td>
<td>3</td>
<td>11.11</td>
<td>34048</td>
<td>7</td>
<td>0.02</td>
<td>2.33</td>
<td>3</td>
</tr>
<tr>
<td>Brazil</td>
<td>23</td>
<td>2</td>
<td>8.70</td>
<td>30033</td>
<td>80</td>
<td>0.27</td>
<td>40.00</td>
<td>78</td>
</tr>
<tr>
<td>Spain</td>
<td>38</td>
<td>3</td>
<td>7.89</td>
<td>70393</td>
<td>8</td>
<td>0.01</td>
<td>2.67</td>
<td>4</td>
</tr>
<tr>
<td>Uruguay</td>
<td>26</td>
<td>2</td>
<td>7.69</td>
<td>31228</td>
<td>22</td>
<td>0.07</td>
<td>11.00</td>
<td>20</td>
</tr>
<tr>
<td>Latvia</td>
<td>27</td>
<td>2</td>
<td>7.41</td>
<td>29919</td>
<td>38</td>
<td>0.13</td>
<td>19.00</td>
<td>36</td>
</tr>
<tr>
<td>Chile</td>
<td>29</td>
<td>2</td>
<td>6.90</td>
<td>37760</td>
<td>6</td>
<td>0.02</td>
<td>3.00</td>
<td>4</td>
</tr>
<tr>
<td>Mexico</td>
<td>30</td>
<td>2</td>
<td>6.67</td>
<td>42819</td>
<td>539</td>
<td>1.26</td>
<td>269.50</td>
<td>537</td>
</tr>
<tr>
<td>Germany-West</td>
<td>30</td>
<td>2</td>
<td>6.67</td>
<td>39433</td>
<td>26</td>
<td>0.07</td>
<td>13.00</td>
<td>24</td>
</tr>
<tr>
<td>France</td>
<td>31</td>
<td>2</td>
<td>6.45</td>
<td>47921</td>
<td>12</td>
<td>0.03</td>
<td>6.00</td>
<td>10</td>
</tr>
<tr>
<td>Slovenia</td>
<td>32</td>
<td>2</td>
<td>6.25</td>
<td>36018</td>
<td>8</td>
<td>0.02</td>
<td>4.00</td>
<td>6</td>
</tr>
<tr>
<td>Hungary</td>
<td>34</td>
<td>2</td>
<td>5.88</td>
<td>38496</td>
<td>4</td>
<td>0.01</td>
<td>2.00</td>
<td>2</td>
</tr>
<tr>
<td>United States</td>
<td>24</td>
<td>1</td>
<td>4.17</td>
<td>34876</td>
<td>528</td>
<td>1.51</td>
<td>528.00</td>
<td>528</td>
</tr>
<tr>
<td>Norway</td>
<td>25</td>
<td>1</td>
<td>4.00</td>
<td>35188</td>
<td>160</td>
<td>0.45</td>
<td>160.00</td>
<td>160</td>
</tr>
<tr>
<td>Italy</td>
<td>28</td>
<td>1</td>
<td>3.57</td>
<td>35264</td>
<td>2</td>
<td>0.01</td>
<td>2.00</td>
<td>2</td>
</tr>
<tr>
<td>Estonia</td>
<td>28</td>
<td>1</td>
<td>3.57</td>
<td>33668</td>
<td>2</td>
<td>0.01</td>
<td>2.00</td>
<td>2</td>
</tr>
<tr>
<td>Slovakia</td>
<td>29</td>
<td>1</td>
<td>3.45</td>
<td>33345</td>
<td>2</td>
<td>0.01</td>
<td>2.00</td>
<td>2</td>
</tr>
<tr>
<td>Sweden</td>
<td>30</td>
<td>1</td>
<td>3.33</td>
<td>37202</td>
<td>18</td>
<td>0.05</td>
<td>18.00</td>
<td>18</td>
</tr>
</tbody>
</table>

Countries without NURs: Columbia (22 surveys), Czech Republic (32), Finland (27), Germany-East (24), Great Britain (30), Lithuania (22), The Netherlands (33), Peru (22), Poland (32), and Switzerland (20)
Table 5: Non-unique Records in Surveys Conducted in Four Periods

<table>
<thead>
<tr>
<th>Period</th>
<th>Number of surveys</th>
<th>Number of surveys with NURs</th>
<th>Percent of surveys with NURs</th>
<th>Number of records</th>
<th>Number of NURs</th>
<th>Percent of NURs</th>
<th>Number of records in surveys with NURs</th>
<th>Percent of NURs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1966-1980a</td>
<td>17</td>
<td>1</td>
<td>5.88</td>
<td>32011</td>
<td>52</td>
<td>0.16</td>
<td>1769</td>
<td>2.94</td>
</tr>
<tr>
<td>1981-1996b</td>
<td>284</td>
<td>30</td>
<td>10.56</td>
<td>351929</td>
<td>2949</td>
<td>0.84</td>
<td>44686</td>
<td>6.60</td>
</tr>
<tr>
<td>1997-2005c</td>
<td>558</td>
<td>70</td>
<td>12.54</td>
<td>707262</td>
<td>1721</td>
<td>0.24</td>
<td>95157</td>
<td>1.81</td>
</tr>
<tr>
<td>2006-2013d</td>
<td>862</td>
<td>61</td>
<td>7.08</td>
<td>1197858</td>
<td>1171</td>
<td>0.10</td>
<td>105115</td>
<td>1.11</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>1721</strong></td>
<td><strong>162</strong></td>
<td><strong>9.41</strong></td>
<td><strong>2289060</strong></td>
<td><strong>5893</strong></td>
<td><strong>0.26</strong></td>
<td><strong>246727</strong></td>
<td><strong>2.39</strong></td>
</tr>
</tbody>
</table>

Countries from the following survey projects / waves are included in the respected periods:

a PA2, PA8NS, PPE7N

6 Implications for statistical analysis

Are rare occurrences of NURs problematic for statistical analyses? The answer to this question depends on the type of estimates of interest. A duplicated extreme value may lead to the recategorization of a case from an outlier to a “regular” case. The resulting inclusion of outliers in research of, for example, the size of largest households in different countries, or their changes over time, may lead to distorted results. In correlation and regression models, a single outlier may significantly influence the results (Treiman, 2009, pp. 94–96), and this is even more likely if the outlier is duplicated. However, what is particularly important for NURs is the pattern of values on all variables taken into account in the analysis. A particular pattern of values in a single duplicate record may constitute a “deviant” case, influencing taxonomic procedures in which respondents are clustered in a multidimensional space.

The statistical effects of a large number of NURs for regression analysis depend on their distribution. If these records are distributed randomly, they artificially increase the significance level of the coefficients but do not affect their values. However, in practice, researchers do not know how these NURs are distributed and what their effect can be.

We examined bivariate correlations \( r_{xy} \) of selected variables (general trust, trust in the parliament, trust in the judiciary, and signing petitions) with a dummy variable identifying NURs in all surveys with NURs. The proportion of correlation significantly different from zero ranges from 21% (signing petitions) to 39% (trust in parliament). For these four selected variables, the maximum value of \( |r_{xy}| \) ranges from 0.12 to 0.22, which shows that NURs cannot be disregarded in more complex analyses. For an assessment of the severity of the bias induced by NURs see Sarracino and Mikucka (2017, in this issue).

7 Discussion and conclusions

Survey methodology is concerned not only with identifying biases and errors that appear in the process of conducting surveys, but also with studying their sources, correlates and consequences (e.g., Alwin, 2007; Andersen, Kasper, Frankel, & Associates, 1979; Brown, 1967; Groves, 1989; Groves & Lyberg, 2010; Weisberg, 2005). This paper focuses on identifying the problem of NURs and describing their distribution in international survey projects. However, a comprehensive program for studying NURs should include a question about the origins of these records. Theoretically, for any pair of identical records there are three possibilities: (a) both records correspond to real respondents, (b) one record corresponds to a real respondent and another one is its duplicate, or (c) both records are fakes.

Based on our probabilistic model, given the parameters of existing surveys in our collection (random samples of heterogeneous populations and a large number of uncorrelated variables), the first possibility (a) is highly unlikely as it would be a miracle (Kruskal, 1988) or improbable coincidence (Diacnosis & Mosteller, 1989). However, it is difficult to exclude the possibility of the natural occurrence of NURs if the simple mono-thematic questionnaire is applied to multi-trait quota samples or samples of homogenous populations (Simmons et al., 2016). For the two remaining possibilities (b and c), one can investigate whether the errors were caused by interviewers, data coders, or data processing staff (e.g., AAPOR, 2003; Crespi, 1945; Koczela et al., 2015; Scheiner, Pennie, & Newbrough, 1988; Winker, Menold, & Porst, 2013).
<table>
<thead>
<tr>
<th>Mode</th>
<th>Number of surveys</th>
<th>Number of NURs</th>
<th>Number of surveys</th>
<th>Number of NURs</th>
<th>Number of surveys</th>
<th>Number of NURs</th>
<th>Number of surveys</th>
<th>Number of NURs</th>
<th>Number of surveys</th>
<th>Number of NURs</th>
<th>Number of surveys</th>
<th>Number of NURs</th>
<th>Number of surveys</th>
<th>Number of NURs</th>
<th>Number of surveys</th>
<th>Number of NURs</th>
<th>Number of surveys</th>
<th>Number of NURs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face to Face</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAPI</td>
<td>23</td>
<td>11973</td>
<td>11973</td>
<td>11973</td>
<td>11973</td>
<td>11973</td>
<td>11973</td>
<td>11973</td>
<td>11973</td>
<td>11973</td>
<td>11973</td>
<td>11973</td>
<td>11973</td>
<td>11973</td>
<td>11973</td>
<td>11973</td>
<td>11973</td>
<td>11973</td>
</tr>
<tr>
<td>CAPI</td>
<td>60</td>
<td>33757</td>
<td>33757</td>
<td>33757</td>
<td>33757</td>
<td>33757</td>
<td>33757</td>
<td>33757</td>
<td>33757</td>
<td>33757</td>
<td>33757</td>
<td>33757</td>
<td>33757</td>
<td>33757</td>
<td>33757</td>
<td>33757</td>
<td>33757</td>
<td>33757</td>
</tr>
<tr>
<td>Mixed: Mail and Web</td>
<td>60</td>
<td>103103</td>
<td>103103</td>
<td>103103</td>
<td>103103</td>
<td>103103</td>
<td>103103</td>
<td>103103</td>
<td>103103</td>
<td>103103</td>
<td>103103</td>
<td>103103</td>
<td>103103</td>
<td>103103</td>
<td>103103</td>
<td>103103</td>
<td>103103</td>
<td>103103</td>
</tr>
<tr>
<td>Web questionnaire</td>
<td></td>
<td>879</td>
<td>879</td>
<td>879</td>
<td>879</td>
<td>879</td>
<td>879</td>
<td>879</td>
<td>879</td>
<td>879</td>
<td>879</td>
<td>879</td>
<td>879</td>
<td>879</td>
<td>879</td>
<td>879</td>
<td>879</td>
<td>879</td>
</tr>
<tr>
<td>Mixed: Mail and Web</td>
<td></td>
<td>143</td>
<td>143</td>
<td>143</td>
<td>143</td>
<td>143</td>
<td>143</td>
<td>143</td>
<td>143</td>
<td>143</td>
<td>143</td>
<td>143</td>
<td>143</td>
<td>143</td>
<td>143</td>
<td>143</td>
<td>143</td>
<td>143</td>
</tr>
<tr>
<td>Face to Face</td>
<td></td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Mixed: Mail and Web</td>
<td></td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
</tr>
</tbody>
</table>

*All NURs from web questionnaire have non-unique counterparts in the mailed back mode.

ISSP 2009, Norway.

Only those surveys were taken into account in which documentation clearly specified the mode of data collection.
Table 7
Non-unique Records According to a Sampling Method Used in National Surveys

<table>
<thead>
<tr>
<th>Sampling method</th>
<th>Number of surveys</th>
<th>Number of surveys with NURs</th>
<th>Percent of surveys with NURs</th>
<th>Number of records</th>
<th>Number of NURs</th>
<th>Percent of NURs</th>
<th>Number of records in surveys with NURs</th>
<th>Percent of records in surveys with NURs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple/stratified random sampling</td>
<td>98</td>
<td>6</td>
<td>6.12</td>
<td>147520</td>
<td>192</td>
<td>0.13</td>
<td>10446</td>
<td>1.84</td>
</tr>
<tr>
<td>Multistage individual register</td>
<td>126</td>
<td>8</td>
<td>6.35</td>
<td>179497</td>
<td>85</td>
<td>0.05</td>
<td>14476</td>
<td>0.59</td>
</tr>
<tr>
<td>Multistage address register</td>
<td>176</td>
<td>8</td>
<td>4.55</td>
<td>279235</td>
<td>426</td>
<td>0.15</td>
<td>14669</td>
<td>2.90</td>
</tr>
<tr>
<td>Multistage random route</td>
<td>420</td>
<td>30</td>
<td>7.14</td>
<td>506812</td>
<td>180</td>
<td>0.04</td>
<td>45241</td>
<td>0.40</td>
</tr>
<tr>
<td>Quota</td>
<td>421</td>
<td>55</td>
<td>13.06</td>
<td>503280</td>
<td>3018</td>
<td>0.60</td>
<td>71311</td>
<td>4.23</td>
</tr>
<tr>
<td>Insufficient information(^a)</td>
<td>480</td>
<td>55</td>
<td>11.46</td>
<td>672716</td>
<td>1992</td>
<td>0.30</td>
<td>90584</td>
<td>2.20</td>
</tr>
<tr>
<td>Total</td>
<td>1721</td>
<td>162</td>
<td>9.41</td>
<td>2289060</td>
<td>5893</td>
<td>0.26</td>
<td>246727</td>
<td>2.39</td>
</tr>
</tbody>
</table>

\(^a\) Includes unspecified information for multistage sampling.
Table 8
Correlation of Unique/Non-unique Records (x) with Selected Variables (y) in the Set of 162 National Surveys

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Selected variables$^a$</th>
<th>General trust</th>
<th>Trust in parliament</th>
<th>Trust in the judiciary</th>
<th>Signing petitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of surveys with a given variable</td>
<td></td>
<td>113</td>
<td>102</td>
<td>90</td>
<td>70</td>
</tr>
<tr>
<td>Number of surveys in which $</td>
<td>r_{xy}</td>
<td>\leq 0.05$$^b$</td>
<td></td>
<td>29</td>
<td>33</td>
</tr>
<tr>
<td>Number of surveys in which $</td>
<td>r_{xy}</td>
<td>\leq 0.10$$^c$</td>
<td></td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Maximum value of $</td>
<td>r_{xy}</td>
<td>$</td>
<td></td>
<td>.22</td>
<td>.15</td>
</tr>
</tbody>
</table>

$^a$ General trust and Signing petition are binary variables while Trust in parliament and Trust in the judiciary have 11-point scales

$p < .05$ for samples with $N_r > 1000$

$p < .005$ for samples with $N_r > 1000$

Some readers may be curious as to why the NURs reported in this paper had not been detected earlier by organizations conducting or archiving surveys. In our view, this is because the duplicated sequences of respondents’ answers are “hidden” among many additional variables (e.g., technical ones) and therefore routine procedures (available in all statistical packages such as SPSS, Stata, and R) are insufficient. In recent research, finding duplicates was limited to small subsets of questionnaire items (Blasius & Thiessen, 2012) or to establishing the likelihood of datasets containing duplicates (Kuriakose & Robbins, 2015).

In this paper we described how NURs are distributed across projects, countries, time, modes of data collection, and sampling methods. Of course, researchers can analyze additional correlates of NURs, such as demographic characteristics of respondents or particular properties of national surveys. Ideally, searches for significant correlates should be motivated by specific hypotheses about where NURs are concentrated. Further analyses into circumstances conducive to the occurrence of NURs may shed light on the mechanisms of their generation.

The presence of NURs has consequences for results of substantive research. As shown, NURs may bias the estimates of statistical models. For further analyses, we suggest treating NURs as a type of measurement error. These errors, shown to be voluminous in some national surveys, need to be controlled for in secondary data analysis, since they reduce confidence in data and their effects potentially distort the results of substantive research. To facilitate analyses of the consequences of NURs we recommend that they be retained in datasets but flagged (by a dummy variable). Such analyses could have implications for already published work using international survey projects with NURs, and future research using these datasets.

The international survey projects used in this paper have been extensively exploited in the past by many researchers. The estimated number of publications relying on these projects’ data differs depending on the source: based on information from the projects’ web pages it is over 11,000, according to Google Scholar – over 25,000, and according to the Web of Science Core Collection – over 2,000 publications and almost 20,000 citations (see Appendix A2). In the spirit of good science, authors may want to consider replication of their analyses with the goal of eliminating NURs or controlling for their presence (King, 1995).

Most of the international survey projects analyzed in this paper are ongoing endeavors. Since the technology of conducting and controlling surveys steadily improves, in the future NURs may disappear altogether. However, the existing NURs should be retained in combined data files of new and old waves. If NURs are flagged, they can be used as controls in cross-time analyses. We provide a complete list of NURs (see footnote 3) for the analyzed national surveys.5

It has not escaped our attention that NURs have multifaceted implications for the study of deviance in social sciences. In particular, NURs reveal malfunction of the infrastructure of scientific research by exposing lapses in controlling the quality of data production. Since, for a long time NURs have been largely neglected, the current interest in their study within the context of deviance in social sciences presents a new challenge.

Acknowledgements

This work was supported by the grant “Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Level Modeling in Cross-National Perspective” from the (Polish) National Science Centre (2012/06/M/HS6/00322). Earlier versions of this paper were presented by Przemek Powałko at “Modes, Measurement, Modelling: Achieving Equivalence in Quanti-

5We informed all providers of data for our project about NURs. We obtained a positive response with regard to retaining NURs from ISSP, ESS, WVS, and AMB.
References


Tomescu-Dubrow, I. & Słomczynski, K. M. (2014). Democratic values and protest behavior: data harmonization,
measurement comparability, and multi-level modeling in cross-national perspective. Research & Methods, 23(1), 103–114.


**Data sources**


European Social Survey (ESS). Waves 1–6. (2002–2013). Members of the project [producers]. Norwegian Social Science Data Services, Norway — Data Archive and distributor of ESS.


## Table A1

### Homepages of 22 International Survey Projects

<table>
<thead>
<tr>
<th>Project</th>
<th>Official name of project</th>
<th>Homepage</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABS</td>
<td>Asian Barometer</td>
<td><a href="http://www.asianbarometer.org">http://www.asianbarometer.org</a></td>
</tr>
<tr>
<td>AFB</td>
<td>Afrobarometer</td>
<td><a href="http://afrobarometer.org">http://afrobarometer.org</a></td>
</tr>
<tr>
<td>AMB</td>
<td>Americas Barometer</td>
<td><a href="http://www.vanderbilt.edu/latpop">http://www.vanderbilt.edu/latpop</a></td>
</tr>
<tr>
<td>ARB</td>
<td>Arab Barometer</td>
<td><a href="http://www.arabbarometer.org">http://www.arabbarometer.org</a></td>
</tr>
<tr>
<td>ASES</td>
<td>Asia Europe Survey</td>
<td><a href="http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/22324?q=asia+europe+survey">http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/22324?q=asia+europe+survey</a></td>
</tr>
<tr>
<td>CB</td>
<td>Caucasus Barometer</td>
<td><a href="http://www.crrccenters.org">http://www.crrccenters.org</a></td>
</tr>
<tr>
<td>EQLS</td>
<td>European Quality of Life Survey</td>
<td><a href="http://discover.ukdataservice.ac.uk/Catalogue/?sn=7348">http://discover.ukdataservice.ac.uk/Catalogue/?sn=7348</a></td>
</tr>
<tr>
<td>ESS</td>
<td>European Social Survey</td>
<td><a href="http://www.europeansocialsurvey.org">http://www.europeansocialsurvey.org</a></td>
</tr>
<tr>
<td>EVS</td>
<td>European Values Study</td>
<td><a href="http://www.europevaluesstudy.eu">http://www.europevaluesstudy.eu</a></td>
</tr>
<tr>
<td>ISSP</td>
<td>International Social Survey Programme</td>
<td><a href="http://www.issp.org">http://www.issp.org</a></td>
</tr>
<tr>
<td>LB</td>
<td>Latinobarometro</td>
<td><a href="http://www.latinobarometro.org">http://www.latinobarometro.org</a></td>
</tr>
<tr>
<td>NBB</td>
<td>New Baltic Barometer</td>
<td><a href="http://discover.ukdataservice.ac.uk/catalogue/?sn=6510">http://discover.ukdataservice.ac.uk/catalogue/?sn=6510</a></td>
</tr>
<tr>
<td>PA8NS</td>
<td>Political Action - An Eight Nation Study</td>
<td><a href="http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/07777">http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/07777</a></td>
</tr>
<tr>
<td>PPE7N</td>
<td>Political Participation and Equality in Seven Nations</td>
<td><a href="http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/07768">http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/07768</a></td>
</tr>
<tr>
<td>VPCPCE</td>
<td>Values and Political Change in Postcommunist Europe</td>
<td><a href="http://discover.ukdataservice.ac.uk/catalogue/?sn=4129">http://discover.ukdataservice.ac.uk/catalogue/?sn=4129</a></td>
</tr>
<tr>
<td>WVS</td>
<td>World Values Survey</td>
<td><a href="http://www.worldvaluessurvey.org">http://www.worldvaluessurvey.org</a></td>
</tr>
</tbody>
</table>

*a For projects that do not have their own web pages, the archiving organization web page was used as a source.*
### Table A2

**Estimated Number of Publications Using Data from International Survey Projects**

<table>
<thead>
<tr>
<th>Project</th>
<th>Number of publications listed in project homepages&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Number of publications found on Google Scholar&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Number of citations in Web of Science&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABS 1</td>
<td>322</td>
<td>177</td>
<td>(354)</td>
</tr>
<tr>
<td>AFB 2</td>
<td>428</td>
<td>1307</td>
<td>(5230)</td>
</tr>
<tr>
<td>AMB3</td>
<td>312</td>
<td>251</td>
<td>(502)</td>
</tr>
<tr>
<td>ARB 4</td>
<td>30</td>
<td>174</td>
<td>(348)</td>
</tr>
<tr>
<td>ASES 5</td>
<td>1</td>
<td>37</td>
<td>(74)</td>
</tr>
<tr>
<td>CB 6</td>
<td>96</td>
<td>66</td>
<td>(164)</td>
</tr>
<tr>
<td>CDCEE 7</td>
<td>1</td>
<td>81</td>
<td>(163)</td>
</tr>
<tr>
<td>CNEP 8</td>
<td>65</td>
<td>49</td>
<td>(326)</td>
</tr>
<tr>
<td>EB 9</td>
<td>825</td>
<td>1167</td>
<td>(40000)</td>
</tr>
<tr>
<td>EQLS 10</td>
<td>70</td>
<td>915</td>
<td>(1830)</td>
</tr>
<tr>
<td>ESS 11</td>
<td>1362</td>
<td>4600</td>
<td>(13800)</td>
</tr>
<tr>
<td>EVS 12</td>
<td>1384</td>
<td>3293</td>
<td>(9878)</td>
</tr>
<tr>
<td>ISJP 13</td>
<td>2</td>
<td>230</td>
<td>(461)</td>
</tr>
<tr>
<td>ISSP 14</td>
<td>6569</td>
<td>1443</td>
<td>(9660)</td>
</tr>
<tr>
<td>LB 15</td>
<td>54</td>
<td>1437</td>
<td>(4600)</td>
</tr>
<tr>
<td>LITS 16</td>
<td>195</td>
<td></td>
<td>(391)</td>
</tr>
<tr>
<td>NBB 17</td>
<td>27</td>
<td>118</td>
<td>(237)</td>
</tr>
<tr>
<td>PA2 18</td>
<td>12</td>
<td>46</td>
<td>(93)</td>
</tr>
<tr>
<td>PA8NS 19</td>
<td>50</td>
<td>78</td>
<td>(156)</td>
</tr>
<tr>
<td>PPE7N 20</td>
<td>8</td>
<td>23</td>
<td>(47)</td>
</tr>
<tr>
<td>VPCPCE 21</td>
<td>30</td>
<td></td>
<td>(60)</td>
</tr>
<tr>
<td>WVS 22</td>
<td>128</td>
<td>9334</td>
<td>(28003)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>11746</strong></td>
<td><strong>25051</strong></td>
<td><strong>(116377)</strong></td>
</tr>
</tbody>
</table>

<sup>a</sup>Data gathered on 2015-02-06.

<sup>b</sup>Data gathered on 2015-03-19. For the total number of items found on Google Scholar for a given project (provided in parentheses), we estimated the number of publications that refer to the project data in two steps: first, we decreased the total number of items proportionally to the number of relevant waves (e.g. for Eurobarometer we took 7 waves out of 80, i.e. 40,000 * 0.0875); second, for large projects with the total number of items over 3000, we divided this number by 3; for the remaining projects we divided this number by 2.

<sup>c</sup>Data gathered on 2015-03-31

The following expressions have been used for searches: