Survey Research Methods (2017)
Vol. 11, No. 1, pp. 1-16
doi:10.18148/srm/2017.v11i1.6557

ISSN 1864-3361
http://www.surveymethods.org

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

Contact information: Kazimierz M. Slomczynski, IFiS-CONSIRT at the Polish Academy of Sciences, Nowy Swiat 72, 00-330 Warsaw, Poland (email: slomczynski.1@osu.edu)

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 the adult population of a given country or territory; (d) the questionnaires contain questions about political attitudes and behaviors; (e) the data are freely available; and (f) survey documentation (study description, codebook and/or questionnaire) is provided in English. This collection of data (see Table 1) covers a large majority of cross-national surveys used in academic research publications (Curtice, 2007; Heath, Fisher, & Smith, 2005; Smith, 2015); it comes from a study of democratic values and political participation (for a detailed description of the datasets, see: Tomescu-Dubrow & Slomczynski, 2014, 2016). Appendix table A1 provides addresses of the survey projects' homepages.[1]

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 (i.e., 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)[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 (Latinobarómetro, 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%). In total, among 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.

[1] Links 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 datafiles-v2.xlsx, pub-3-_README.docx, pub-3-correcting and converting files.docx, pub-3-merging ABS.docx, pub-3-merging and patching EVS and WVS.docx, pub-3-patching EB.docx, pub-3-patching ESS.docx, pub-3-patching 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).

[2] For 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
Basic Characteristics of 22 International Survey Projects

Survey project[a]	Time span	Number of surveys	Number of distinct countries[b]	Average number of questions	Average sample size	Number of records
ABS	2001-2011	30	13	174	1456	43691
AFB	1999-2009	66	20	210	1499	98942
AMB	2004-2012	92	24	178	1645	151341
ARB	2006-2011	16	11	219	1230	19684
ASES	2000	18	18	193	1014	18253
CB	2009-2012	12	3	275	2052	24621
CDCEE	1990-2001	27	16	299	1071	28926
CNEP[cd]	2004-2006	8	8	294	1672	13372
EB[c]	1983-2012	152	37	342	913	138753
EQLS	2003-2012	93	35	167	1135	105527
ESS	2002-2013	146	32	223	1928	281496
EVS	1981-2009	128	50	347	1301	166502
ISJP	1991-1997	21	14	205	1229	25805
ISSP[c]	1985-2013	363	53	88	1359	493243
LB	1995-2010	260	19	251	1134	294965
LITS	2006-2010	64	35	636	1060	67866
NBB	1993-2004	18	3	172	1200	21601
PA2[d]	1979-1981	3	3	271	1352	4057
PA8NS	1973-1976	8	8	345	1574	12588
PPE7N	1966-1971	7	7	299	2360	16522
VPCPCE[d]	1993	5	5	193	945	4723
WVS	1981-2008	184	89	221	1394	256582
Total	1966-2013	1721	142	228	1330	2289060

[a] Data were downloaded at the turn of 2013/2014. For detailed dates and links to data sources, see supplementary materials.
[b] Countries or territories.
[c] For CNEP, EB, and ISSP, only selected survey editions were used.
[d] 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

Survey project	Number of countries with NURs	Percent of countries with NURs	surveys with NURs	surveys with NURs	Number of NURs	Percent of NURs	Number of records in surveys with NURs	Percent of NURs in surveys with NURs
LB	13	68.42	32	12.31	1225	0.42	35633	3.44
AMB	10	41.67	12	13.04	48	0.03	22431	0.21
ISSP	19	35.85	31	8.54	923	0.19	59587	1.55
WVS	31	34.83	36	19.57	1970	0.77	54449	3.62
CB	1	33.33	1	8.33	2	0.01	1975	0.10
NBB	1	33.33	1	5.56	2	0.01	1987	0.10
ABS	3	23.08	3	10.00	12	0.03	7289	0.16
EB	8	21.62	11	7.24	797	0.57	10773	7.40
LITS	7	20.00	7	10.94	32	0.05	7001	0.46
EQLS	7	20.00	8	8.60	40	0.04	8549	0.47
AFB	4	20.00	4	6.06	28	0.03	9092	0.31
CDCEE	3	18.75	3	11.11	168	0.58	3740	4.49
ESS	5	15.63	5	3.42	14	0.00	10227	0.14
PPE7N	1	14.29	1	14.29	52	0.31	1769	2.94
EVS	5	10.00	5	3.91	570	0.34	10224	5.58
ISJP	1	7.14	1	4.76	2	0.01	1001	0.20
ASES	1	5.56	1	5.56	8	0.04	1000	0.80
Total	80	56.34	162	9.41	5893	0.26	246727	2.39

questions.

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

$$p \approx 1 - \exp \frac{-N_r^2}{2^{k+1}} \quad .$$

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

$$N_r \approx \sqrt{-2^{k+1} \log (1 - p)} \quad .$$

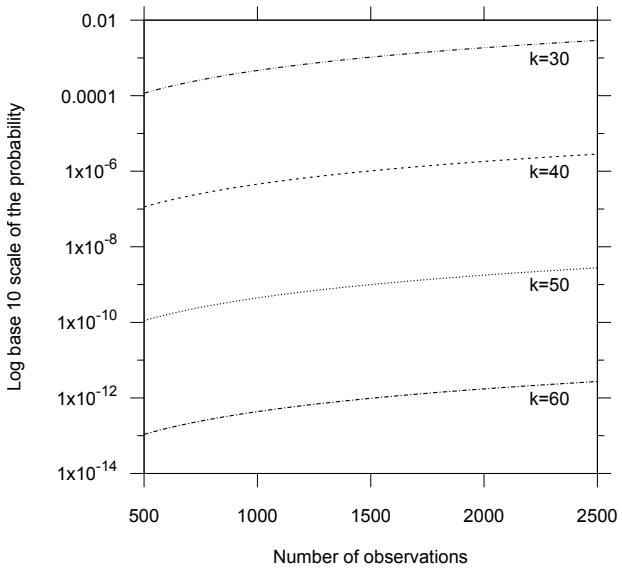

Figure 1. Estimated Probability of at least One Duplicate by Sample Size and Number of Independent Binary Variables (k)

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.

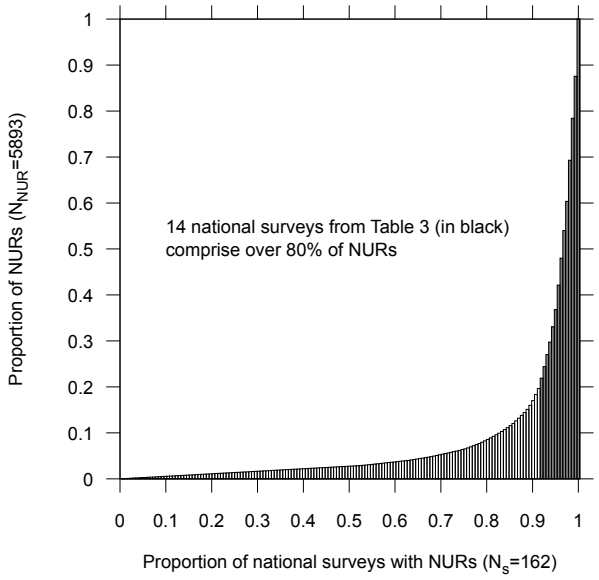

Figure 2. Lorenz Curve for Non-unique Records in 162 National Surveys from 17 Projects Shown in Table 2

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[4] – Algeria, Bangladesh, Ethiopia, Iraq, Kyrgyzstan, Rwanda, Saudi Arabia, and Tajikistan – 9 out of 17 surveys have NURs, indicating a high proportion of affected surveys. In these countries the infrastructure for conducting social surveys is not firmly established.

In contrast, in Table 4 we include countries with more than 20 surveys, and provide a detailed analysis of the distribution of NURs. In this group of countries, ten do not have any NURs. In the remaining 27 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

[4]Here 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

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

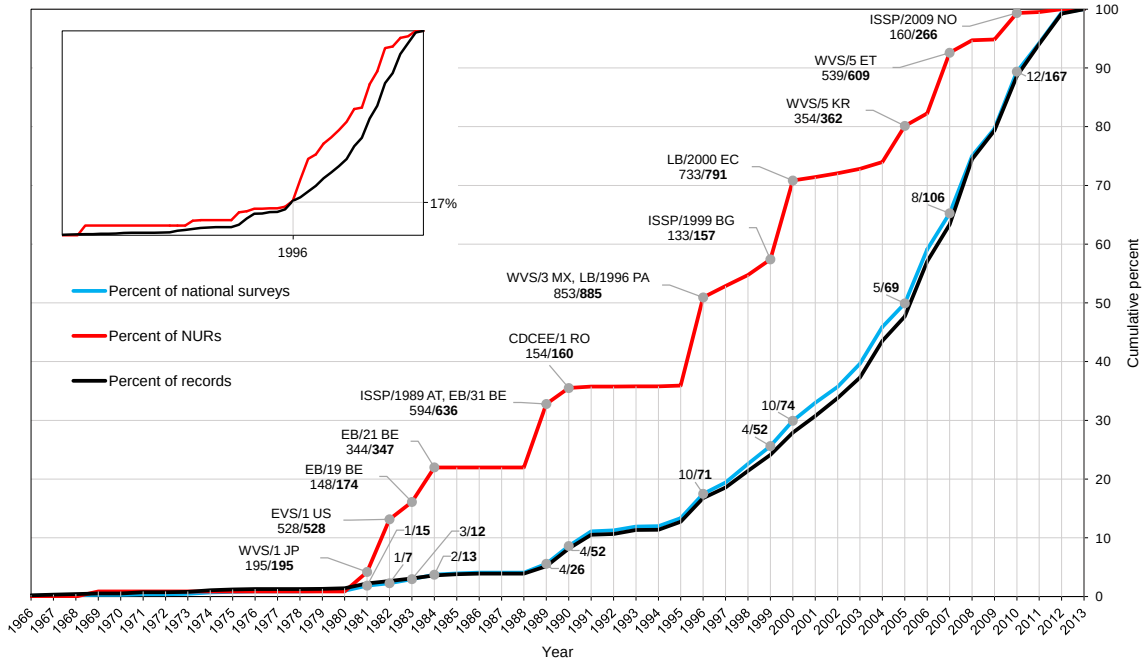

Figure 3. Cumulative Percent Distribution of National Surveys, Records and NURs. Percent of national surveys (blue line) covers 1,721 surveys. The marked years here are those in which the largest increase of NURs occurred. The first number is the number of surveys with NURs in the given year; the second 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 surveyâĂŹs 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 NURsâĂŹ 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.

Table 4
Non-unique Records in Countries with At Least 20 Surveys, Ordered by the Percent of Surveys with NURs

Country	Number of surveys	Number of surveys with NURS	Percent of surveys with NURs	Number of records	Number of NURs	Percent of NURs	Average number of NURs per survey with NURs	Maximal number of NURs in a survey
Portugal	25	6	24.00	35700	74	0.21	12.33	40
Bulgaria	28	6	21.43	34384	146	0.42	24.33	133
Belgium	20	4	20.00	28400	714	2.51	178.50	344
Guatemala	20	3	15.00	22755	24	0.11	8.00	20
El Salvador	20	3	15.00	23234	6	0.03	2.00	2
Ireland	30	4	13.33	39551	20	0.05	5.00	8
Venezuela	23	3	13.04	28185	70	0.25	23.33	60
Austria	24	3	12.50	31923	430	1.35	143.33	374
Argentina	24	3	12.50	28769	32	0.11	10.67	28
Russia	25	3	12.00	46871	10	0.02	3.33	4
Denmark	27	3	11.11	34048	7	0.02	2.33	3
Brazil	23	2	8.70	30033	80	0.27	40.00	78
Spain	38	3	7.89	70393	8	0.01	2.67	4
Uruguay	26	2	7.69	31228	22	0.07	11.00	20
Latvia	27	2	7.41	29919	38	0.13	19.00	36
Chile	29	2	6.90	37760	6	0.02	3.00	4
Mexico	30	2	6.67	42819	539	1.26	269.50	537
Germany-West	30	2	6.67	39433	26	0.07	13.00	24
France	31	2	6.45	47921	12	0.03	6.00	10
Slovenia	32	2	6.25	36018	8	0.02	4.00	6
Hungary	34	2	5.88	38496	4	0.01	2.00	2
United States	24	1	4.17	34876	528	1.51	528.00	528
Norway	25	1	4.00	35188	160	0.45	160.00	160
Italy	28	1	3.57	35264	2	0.01	2.00	2
Estonia	28	1	3.57	33668	2	0.01	2.00	2
Slovakia	29	1	3.45	33345	2	0.01	2.00	2
Sweden	30	1	3.33	37202	18	0.05	18.00	18

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)

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 – 0.30%. 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 percent 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 dace of multistage address register reaching almost 3%.

Table 5
Non-unique Records in Surveys Conducted in Four Periods

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

Countries from the following survey projects / waves are included in the respected periods:
[a] PA2, PA8NS, PPE7N
[b] CDCEE/1, EB/19-31, EVS/1-2, ISJP, ISSP/1985-1996, LB/1995-1996, NBB/1-3, PA2, VPCPCE, WVS/1-3
[c] ABS/1-2, AFB/1-3, AMB/2004, ASES, CDCEE/2, CNEP/3, EB/54.1-62.2, EQLS/1, ESS/1-2, EVS/3, ISSP/1996-2006, LB/1997-2005, NBB/4-6, WVS/3-5
[d] ABS/2-3, AFB/3-4, AMB/2006-2012, ARB/1-2, CB/2009-2012, CNEP/3, EB/73.4-77.3, EQLS/2-3, ESS/2-6, EVS/4, ISSP/2004-2011, LB/2006-2010, LITS/1-2, WVS/5

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 reclassification 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 (Diaconis & 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; Schreiner, Pennie, & Newbrough, 1988; Winker, Menold, & Porst, 2013).

Table 6
Non-unique Records According to the Mode of Data Collection in ISSP, LB, and WVS

Mode	Number of surveys[a]	Number of surveys with NURs	Percent of surveys with NURs	Number of records	Number of NURs	Percent of NURs	Number of records in surveys with NURs	Percent of NURs in surveys with NURs	Maximal number of NURs in a survey
Face to Face									
Not specified	444	70	15.77	550053	3702	0.67	93347	3.97	733
PAPI	124	12	9.68	173853	102	0.06	27735	0.37	60
CAPI	28	2	7.14	35357	22	0.06	3307	0.67	20
Self-Completion									
Mailed back	70	6	8.57	103103	36	0.03	11973	0.30	23
Pickup	13	0	0.00	15937	0	0.00	0	-	0
Mixed: Mail and Web[b]									
Mailed back	1	1	100.00	879	143	10.99	1456	10.99	160
Web questionaire				577	17[c]				

[a] Only those surveys were taken into account in which documentation clearly specified the mode of data collection.
[b] ISSP 2009, Norway.
[c] All NURs from web questionnaire have non-unique counterparts in the mailed back mode.

Table 7
Non-unique Records According to a Sampling Method Used in National Surveys

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

[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

	Selected variables[a]			
Characteristics	General trust	Trust in parliament	Trust in the judiciary	Signing petitions
Number of surveys with a given variable	113	102	90	70
Number of surveys in which $\|r_{xy}\| \leq 0.05$[b]	29	33	32	12
Number of surveys in which $\|r_{xy}\| \leq 0.10$[c]	3	7	3	3
Maximum value of $\|r_{xy}\|$	.22	.15	.13	.12

[a] General trust and Signing petition are binary variables while Trust in parliament and Trust in the judiciary have 11-point scales
[b] $p < .05$ for samples with $N_r > 1000$
[c] $p < .005$ 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-

[5] We 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.

tative Research", ESA RN21/EQMC conference (GESIS, Mannheim, October 24-25, 2014) and the 6th Conference of the European Survey Research Association (Reykjavik, July 13-17, 2015). We thank Elizabeth Zechmeister and Mitchell Seligson (Vanderbilt University), Marta Kołczyńska (The Ohio State University), Irina Tomescu-Dubrow (Polish Academy of Sciences), Markus Quandt (GESIS), Dominique Joye (Université de Lausanne), Jörg Blasius (Universität Bonn), the SRM editor and two anonymous reviewers for useful comments and remarks.

References

AAPOR. (2003). Interviewer falsification in survey research: current best methods for prevention, detection and repair of its effects. American Association for Public Opinion Research. Retrieved from http://www.amstat.org/sections/srms/falsification.pdf

Alwin, D. F. (2007). *Margins of error: a study of reliability in survey measurement*. New Jersey: John Wiley & Sons.

Andersen, R., Kasper, J., Frankel, M. R., & Associates. (1979). *Total survey error*. San Francisco: Jossey-Bass.

Biemer, P. & Lyberg, L. E. (2003). *Introduction to survey quality*. New Jersey: John Wiley & Sons.

Blasius, J. & Thiessen, V. (2012). *Assessing the quality of survey data*. London: Sage.

Blasius, J. & Thiessen, V. (2015). Should we trust survey data? Assessing response simplification and data fabrication. *Social Science Research*, *52*, 479–493.

Brown, R. V. (1967). Evaluation of total survey error. *Statistician*, *17(4)*, 335–356.

Crespi, L. P. (1945). The cheater problem in polling. *Public Opinion Quarterly*, *9(4)*, 431–445.

Curtice, J. (2007). Comparative opinion surveys. In R. J. Dalton & H.-D. Klingemann (Eds.), *The Oxford handbook of political behavior* (pp. 896–909). Oxford: Oxford University Press.

Diaconis, P. & Mosteller, F. (1989). Method of studying coincidences. *Journal of the American Statistical Association*, *84(408*, 853–861.

Feller, W. (1968). *An introduction to probability theory and its applications (3rd ed., vol.1*. New York: John Wiley & Sons.

Gideon, L. ((2012). *Handbook of survey methodology for the social sciences*. New York: Springer.

Groves, R. M. (1989). *Survey errors and survey costs*. New Jersey: John Wiley & Sons.

Groves, R. M. & Lyberg, L. E. (2010). Total survey error: past, present, and future. *Public Opinion Quarterly*, *74(5)*, 849–879.

Harkness, J. A., van de Vijver, F. J. R., & Mohler, P. P. ((2003). *Cross-cultural survey methods*. New Jersey: John Wiley & Sons.

Heath, A., Fisher, S., & Smith, S. (2005). The globalization of public opinion research. *Annual Review of Political Science*, *8(1)*, 297–333.

ISSP. (2009). International Social Survey Programme: leisure time and sports – ISSP 2007. GESIS Data Archive, Cologne. ZA4850 Data file Version 2.0.0. doi:10.4232/1.10079

King, G. (1995). Replication, replication. *PS: Political Science & Politics*, *28(3)*, 444–452.

Koczela, S., Furlong, C., McCarthy, J., & Mushtaq, A. (2015). Curbstoning and beyond: confronting data fabrication in survey research. *Statistical Journal of the IAOS*, *31(3)*, 413–422.

Kohler, U. (2007). Surveys from inside: an assessment of unit nonresponse bias with internal criteria. *Survey Research Methods*, *1(2)*, 55–67.

Kruskal, J. (1988). Miracles and statistics: the casual assumption of independence. *Journal of the American Statistical Association*, *83(404)*, 929–940.

Kuriakose, N. & Robbins, M. (2015). Don't get duped: fraud through duplication in public opinion surveys. Retrieved from http://ssrn.com/abstract=2580502

Latinobarómetro. (2000). Latinobarómetro, Waves 1995-1998, 2000-2010. Merged datasets. Retrieved from http://www.latinobarometro.org/

Lyberg, L. E., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., & Trewin, D. (1997). *Survey measurement and process quality*. New Jersey: John Wiley & Sons.

McNabb, D. E. (2014). *Nonsampling error in social surveys*. Thousand Oaks, CA: Sage.

Sarracino, F. & Mikucka, M. (2017). Estimation bias due to duplicated observations: a monte carlo simulation. *Survey Research Methods*, *11*(1), 17–44. doi:10.18148/srm/2017.v11i1.7149

Schreiner, I., Pennie, K., & Newbrough, J. (1988). Interviewer falsification in census bureau surveys. In *Proceedings of the american statistical association (survey research methods section)* (pp. 491–496).

Simmons, K., Mercer, A., Schwarzer, S., & Kennedy, C. (2016). Evaluating a new proposal for detecting data falsification in surveys. Retrieved from http://www.pewresearch.org/2016/02/23/evaluating-a-new-proposal-for-detecting-data-falsification-in-surveys

Smith, T. W. (2015). Resources for conducting cross-national survey research. *Public Opinion Quarterly*, *79*(S1), 404–409.

Survey Research Center, Institute for Social Research, University of Michigan. (2010). Guidelines for best practice in cross-cultural surveys. Retrieved from http://www.ccsg.isr.umich.edu

Tomescu-Dubrow, I. & Slomczynski, 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.

Tomescu-Dubrow, I. & Slomczynski, K. M. (2016). Harmonization of cross-national survey projects on political behavior: developing the analytic framework of survey data recycling. *International Journal of Sociology*, *46(1)*, 58–72.

Treiman, D. J. (2009). *Quantitative data analysis. Doing social research to test ideas*. San Francisco: Jossey-Bass.

Weisberg, H. F. (2005). *The total survey error approach: a guide to the new science of survey research*. Chicago: University of Chicago Press.

Winker, P., Menold, N., & Porst, R. (Eds.). (2013). *Interviewers' deviations in surveys – impact, reasons, detection and prevention*. New York: Peter Lang, PL Academic Research.

Data sources

Asian Barometer (ABS). Waves 1–3. (2001–2011). Members of the project [producers]. Asian Barometer Project [distributor].

Afrobarometer (AFB). Waves 1–4. (1999–2009). Afrobarometer Network [producer and distributor].

Americas Barometer (AMB). Waves 1–5. (2004–2012). Latin American Public Opinion Project (LAPOP) [producer and distributor].

Arab Barometer (ARB). Waves 1 & 2. (2006 & 2011). Tessler, M., Jamal, A., Bedaida, A., et al. [producers, Wave 1], Jamal, A., Tessler, M., Shikaki, K., et al [producers, Wave 2]. Arab Democracy Barometer and Arab Reform Initiative [distributor].

Asia Europe Survey (ASES). Wave 1. (2000). Takashi, I. [producer]. Inter–University Consortium of Political and Social Research [distributor].

Caucasus Barometer (CB), Waves 1–4. (2009–2012). The Caucasus Research Resource Centers [producer and distributor].

Consolidation of Democracy in Central and Eastern Europe (CDCEE). Waves 1 & 2. (1990 & 2001). Rotman, D., Raychev, A., Stoychev, K., Hartl, J. Misovic, J., Mansfeldová, Z., at al. [producers]. GESIS Data Archive [distributor].

Comparative National Election Project (CNEP). Wave 3 (2006). Members of the project [producers]. Mershon Center for International Security Studies [distributor].

Eurobarometer (EB). Selected 7 Waves. (1983–2012). European Commission [producer]. GESIS Data Archive [distributor].

European Quality of Life Survey (EQLS). Waves 1–3. (2003–2012). European Foundation for the Improvement of Living and Working Conditions [producer and distributor].

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

European Values Study (EVS). Waves 1–4. (1981–2009). European Values Study Foundation [producer]. GESIS Data Archive [distributor].

International Social Justice Project (ISJP). Waves 1 and 2. (1991–1996). Wegener, B. & Mason, D. [producers]. Inter–University for Political and Social Research [distributor].

International Social Survey Programme (ISSP). Selected 13 Waves. (1985–2013). ISSP Research Group [producer]. GESIS Data Archive [distributor].

Latinobarómetro (LB). Waves 1–15. (1995–2010). Corporatión Latinobarómetro [producer and distributor].

Life in Transition Survey (LITS). Wave 1 & 2. (2006 & 2010). European Bank for Reconstruction and Development [producer and distributor].

New Baltic Barometer (NBB). Waves 1–6. (1993–2004). Rose, R. [producer]. UK Data Service [distributor] Political Action II (PA2). Wave 1. (1981). Allerbeck, K. R., Barnes, S. H., van Deth, J. W. Farah, B. G., Heunks, F. J., Inglehart, R. et al. [producers]. GESIS Data Archive [distributor].

Political Action: An Eight Nation Study (PA8NS). Wave 1. (1976). Barnes, S. H. & Kaase, M. [producers]. Inter–University Consortium for Political and Social Research [distributor].

Political Participation and Equality in Seven Nations (PPE7N). Wave 1. (1971). Verba, S., Nie, N. H., & Kim, J–O. [producers]. Inter–University Consortium for Political and Social Research [distributor].

Values and Political Change in Post–Communist Europe (VPCPCE). Wave 1. (1993). Miller, W.L., White, S., & Heywood, P. [producers]. UK Data Service [distributor].

World Values Survey (WVS). Waves 1–5. (1981–2008). Members of the project [producers]. World Values Survey Association [distributor].

Appendix
Tables

Table A1
Homepages of 22 International Survey Projects[a]

Project	Official name of project	Homepage
ABS	Asian Barometer	http://www.asianbarometer.org
AFB	Afrobarometer	http://afrobarometer.org
AMB	Americas Barometer	http://www.vanderbilt.edu/lapop
ARB	Arab Barometer	http://www.arabbarometer.org
ASES	Asia Europe Survey	http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/22324?q=asia+europe+survey
CB	Caucasus Barometer	http://www.crrccenters.org
CDCEE	Consolidation of Democracy in Central and Eastern Europe	https://dbk.gesis.org/dbksearch/sdesc2.asp?no=4054
CNEP	Comparative National Elections Project	http://www.cnep.ics.ul.pt
EB	Eurobarometer	http://zacat.gesis.org/webview/main.jsp?object=http://zacat.gesis.org/obj/fCatalog/Catalog57
EQLS	European Quality of Life Survey	http://discover.ukdataservice.ac.uk/Catalogue/?sn=7348
ESS	European Social Survey	http://www.europeansocialsurvey.org
EVS	European Values Study	http://www.europeanvaluesstudy.eu
ISJP	International Social Justice Project	https://dbk.gesis.org/dbksearch/sdesc2.asp?no=3522
ISSP	International Social Survey Programme	http://www.issp.org
LB	Latinobarometro	http://www.latinobarometro.org
LITS	Life in Transition Survey	http://www.ebrd.com/what-we-do/economic-research-and-data/data/lits.html
NBB	New Baltic Barometer	http://discover.ukdataservice.ac.uk/catalogue/?sn=6510
PA2	Political Action II	https://dbk.gesis.org/dbksearch/sdesc2.asp?no=1188
PA8NS	Political Action - An Eight Nation Study	http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/07777
PPE7N	Political Participation and Equality in Seven Nations	http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/07768
VPCPCE	Values and Political Change in Postcommunist Europe	http://discover.ukdataservice.ac.uk/catalogue/?sn=4129
WVS	World Values Survey	http://www.worldvaluessurvey.org

[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

	Number of publications listed in				
Project	homepages[a]	Google Scholar[b]		Web of Science[c]	Number of citations in Web of Science[c]
ABS 1	322	177	(354)	4	2
AFB 2	428	1307	(5230)	55	204
AMB3	312	251	(502)	13	27
ARB 4	30	174	(348)	3	6
ASES 5	1	37	(74)	2	0
CB 6	96	66	(164)	0	0
CDCEE 7	1	81	(163)	0	0
CNEP 8	65	49	(326)	3	1
EB 9	825	1167	(40000)	409	4992
EQLS 10	70	915	(1830)	27	116
ESS 11	1362	4600	(13800)	590	3637
EVS 12	1384	3293	(9878)	175	1397
ISJP 13	2	230	(461)	20	518
ISSP 14	6569	1443	(9660)	283	3281
LB 15	54	1437	(4600)	21	156
LITS 16		195	(391)	7	1
NBB 17	27	118	(237)	2	3
PA2 18	12	46	(93)	0	0
PA8NS 19	50	78	(156)	0	0
PPE7N 20	8	23	(47)	0	0
VPCPCE 21		30	(60)	1	0
WVS 22	128	9334	(28003)	472	5385
Total	11746	25051	(116377)	2087	19726

[a] Data gathered on 2015-02-06.
[b] 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.
[c] Data gathered on 2015-03-31

The following expressions have been used for searches:
1 “asian barometer survey”, 2 “afrobarometer” OR “afro-barometer” OR “afro barometer”, 3 “americas barometer”, 4 “arab barometer”, 5 “asia europe survey”, 6 “caucasus barometer”, 7 “consolidation of democracy in central and eastern europe”, 8 “comparative national elections project” OR “comparative national election project”, 9 “eurobarometer”, 10 “european quality of life survey”, 11 “european social survey”, 12 “european values study” OR “european value study” OR “european values survey” OR “european value survey”, 13 “international social justice project”, 14 “international social survey programme” OR “international social survey program”, 15 “latinobarometro” OR “latino barometro” OR “latino barometer” OR “latinobarometer”, 16 “life in transition survey”, 17 “new baltic barometer”, 18 “political action ii”, 19 “political action” “eight nation study”, 20 “political participation and equality” “verba”, 21 “values and political change in post communist europe”, 22 “world values survey” OR “world value survey” OR “world values study” OR “world value study”