CONSIRT Cross-National Studies:
Interdisciplinary Research and Training Program

Working Papers Series
CONSIRT Labs: Methodology of Survey Data Harmonization

11 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
slomczynski.1@osu.edu
Przemek Powałko,
Polish Academy of Sciences, Warsaw, Poland
Tadeusz Krauze,
Hofstra University, Hempstead, NY, USA

All authors contributed equally to this work

CONSIRT, The Ohio State University and the Polish Academy of Sciences
February 2016

Mission

What CONSIRT Working Papers Are

Citation of CONSIRT Working Papers

Copyright Statement

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

October 21, 2015

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*

All authors contributed equally to this work.

Correspondence: slomczynski.1@osu.edu

We report on the existence of non-unique responses in a large collection of social science survey projects. We analyzed 1,721 national surveys in 22 projects, covering 142 countries and 2.3 million respondents, and found a total of 5,893 non-unique records in 162 national surveys from 17 projects in 80 countries. The probability of any non-unique record in an average survey sample is exceedingly small, and although non-unique records constitute a minor fraction of all records, it is unlikely that they are solely the result of random chance. Non-unique records diminish data quality and potentially have undesirable effects on the results of statistical analyses. Identifying non-unique records 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.

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., Lyberg at al., 1997; Biemer & Lyberg, 2003; Harkness, van de Vijver, & Mohler, 2003; Gideon, 2012; 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 *Guidelines for Best Practice in Cross-Cultural Surveys*, 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,” no systematic assessment of the prevalence of these errors has yet been made. Extant research on duplicate records deals with a limited number of variables (Blasius & Thiessen, 2012) or "near-duplicates" (Kuriakose & Robbins, 2015).

We argue that a record is erroneous, or at least suspicious, if it is not unique, that is, when the set of all answers by a given respondent is identical to that of another respondent. In the literature, such records are known as duplicates (e.g., Kuriakose & Robbins, 2015; Blasius & Thiessen, 2015; Koczela, Furlong, McCarthy, & Mushtaq, 2015). This concept may be misleading because it suggests that there is an original that is 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 a *non-unique record* and refer to a "duplicate record" as its synonym.

The commonly used organization of survey data, with cases in rows and variables in columns, facilitates comparing the records of interviewees' responses. The number of rows equals the sample size and the number of columns equals the number of variables included in a given survey. Thus the search for non-unique cases involves the pairwise comparison of all rows for a given set of columns.

Before presenting how frequently non-unique records occur in international survey projects, we describe our database and the method of the performed analysis. The probabilistic model precedes the sections in which we present our findings and illustrate the potential effects of non-unique records on the results of statistical analyses. The paper ends with a summary and discussion.

Data and method

We analyzed a collection of 1,721 national surveys in 22 projects covering 142 countries and 2.3 million respondents. This collection comes from a study of political participation (Tomescu-Dubrow & Slomczynski, 2014). The 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) questionnaires contain questions about political attitudes and behaviors; (e) data are freely available; and (f) survey documentation (study description, codebook and/or questionnaire) is provided in English. The analyzed survey projects (see Table 1) constitute a large majority of all projects satisfying these criteria (Heath, Fisher, & Smith, 2005; Curtice, 2007; Smith, 2015). Appendix A provides addresses of the survey projects' homepages.[1]

□ Table 1 about here.

In order to obtain records of variables corresponding to questionnaire items, that is, 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 non-unique records consisted of pairwise comparisons of each case with every other one within a given national survey. For an average sample size (N=1330) the number of comparisons is close to one million. Response options among the considered variables ranged from dichotomous to hundreds of categories, and comparisons were done on raw values of all

[1] Links to the used source data files are available at a dataverse at https://dataverse.harvard.edu/dataverse/duprecords

variables. This task for all national surveys was done in a relational database with the use of SQL.

Probabilistic model

Are non-unique records for respondents' answers undoubtedly suspicious? The answer to this query depends on the probability of at least two respondents providing the same answers to all questions. For a given number of questions, this probability is determined by the number of respondents, the number of response categories, and the dependence among answers to different questions. In calculating the probability of duplicates, we consider these three factors referring to specific survey projects.

The number of questions addressed to respondend in the survey questionnaires ranges from 88 to 636, with an average of 228 (see Table 1). To estimate the probability of duplicate records, we assume dichotomous answers (binary choices) and the statistical independence of answers to 1/3 of the questions. 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 1/3 of items occur only rarely[2].

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.

Applying the modified birthday problem model to the data in Table 1, an appropriate calculation shows that, for example, for 76 independent binary variables (1/3 of the average number of questions per national survey, i.e., 228) one would need $3.90*10^{10}$ respondents in order to find a pair of identical sets of answers with the probability 0.01. In the case of the lower and upper bounds of the number of questionnaire items (88 and 636), the numbers of respondents needed for a duplicate are respectively 3,285 and $1.15*10^{31}$ (with the same probability 0.01). In other words, even for such a small number of cases as 3,285, we still need 100 samples of this size to expect a single duplicate. The intuitive understanding of the model can be based on the fact that the order of magnitude of the number of respondents ($N \approx 10^3$) is much smaller than the order of magnitude of possible response patterns ($N \approx 10^{10}$ for the average number of questions per national survey, i.e., 228). If duplicates occur in surveys with an average sample size per survey project ranging from 913 to 2,360 respondents and data administrators do not publically comment on their occurrence, such records are undoubtedly suspicious.

Findings

In Table 2 we list international survey projects in which we found non-unique records, with an overall total of 5,893[3]. These records are present in 162 national surveys (9.4% of the total),

[2] This empirical evidence gives only plausible support for our assumption since even zero-correlations do not imply statistical independence.

[3] Among the non-unique records, only 67 are clearly lacking the respondents' answers as if the relevant interviews had been interrupted or not even begun. For the complete list of suspicious records see pub-5-IDs of duplicates.xlsx in a dataverse at http://dx.doi.org/10.7910/DVN/HPXFA1

unequally distributed across survey projects. For example, non-unique records appear in 19.6% of surveys of the World Values Survey (the highest value) and 3.4% of surveys of the European Social Survey (the lowest value). Within each project, there are differences with respect to the number of countries in which surveys have non-unique records. In the extreme case, surveys in 13 out of 19 countries included in Latinobarometro contain non-unique records; for other projects see Table 2. Overall, national surveys in 80 out of 142 countries have non-unique records. Duplicates were found in countries at all levels of economic (e.g., Japan, Mexico, and Ethiopia) and political (e.g., Norway, Romania, and Panama) development. Generally, these results, with numbers defying the odds, show that suspicious records are common and universal.

□ Table 2 about here.

In 52% of the affected surveys a single duplicate record was found. In the remaining 48% we found several patterns of non-unique records, such as multiple doublets or records repeated three, four, or even more times, often in combination. For example, in a survey conducted in Ecuador (Latinobarometro, 2000), 733 non-unique records (i.e., 272 doublets and 63 triplets) are present in the sample of 1,200, which means that over 60% of records are suspicious. In another survey in Norway (International Social Survey Programme, 2009) there are 54 records 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 suspicious records in the sample of 1,456 (11.0%).

The distribution of non-unique records is provided in Figure 1. The share of non-unique records among national surveys is very uneven: 148 surveys contain 20% of all duplicates and the remaining 80% are present in just 14 surveys. Commenting on the first group of surveys, we note that the proportion of non-unique records in the respective samples usually does not exceed 1%. However, the latter group of surveys, with a much larger number of duplicates, we are compelled to treat with suspicion.

□ Figure 1 about here.

In our view, every pair of identical records should be subjected to data quality control. Of course, a large number of duplicates in a particular survey are especially troublesome. Our analysis shows that a large number of non-unique records occurs in the following projects: Consolidation of Democracy in Central and Eastern Europe, Eurobarometer, European Values Study, International Social Survey Programme, Latinobarometro, and World Values Survey. In 14 national surveys, included in these projects, there are more than 10% of non-unique records (see Table 3).

□ Table 3 about here.

Implications

Are rare occurrences troublesome for statistical analyses? The answer to this question depends on the kind of estimates of interest. For example, assume that a researcher is looking for the number of people living in the largest households in different countries. Non-unique records giving the misleadingly high maximum number, strong outliers, may change the placement of this country among others with respect to the analyzed variable. A single outlier may significantly influence the results of correlation and regression models (Treiman, 2009, pp. 94-96), and we note that this is even more so if the outlier is duplicated. However, what is particularly important for non-unique records is the pattern of values on all variables taken into account in the analysis. A particular pattern of a single duplicate record may constitute a “deviant” case, influencing some taxonomic procedures in which respondents are clustered in multidimensional space.

The statistical effects of a large number of non-unique records for regression analysis depend on their distribution. If these records are distributed randomly, removing them decreases the significance level of the coefficients but not their values. However, in practice, researchers do not know how these non-unique records are distributed and what their effect could be. We analyzed the extreme cases, that is, the samples with the highest proportions of non-unique records, as shown in Table 3. For exploratory purposes, we ran a simple linear regression of interest in politics on gender, age, education, and respondent’s reaction to the interview, for samples with included and excluded non-unique records. For illustration, in Table 4 we demonstrate that for education and respondent’s reaction to the interview both the regression coefficients (B, beta) as well as their significance differ depending on the inclusion or exclusion of non-unique records. For an assessment of the severity of the bias induced by non-unique records see Sarracino & Mikucka, forthcoming.

- Table 4 about here.

Discussion and conclusion

This study has implications for (a) already published work using international survey projects with non-unique records, (b) future research using these data sets, and (c) for survey methodology. The estimated number of publications relying on data from analyzed projects 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 2000 publications and almost 20,000 citations (see Appendix B). In the spirit of good science, authors may want to consider replication of their analyses with the goal of eliminating non-unique records or controlling for their presence (King, 1995).

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, and (c) both records are fakes. The first possibility, as a miracle (Kruskal, 1988) or improbable coincidence (Diaconis & Mosteller, 1989), should be rejected on statistical grounds. For the two remaining possibilities, one could investigate whether the errors were caused by interviewers, data coders, or data processing staff (Crespi, 1945; Schreiner, Pennie, & Newbrough, 1988; American Association for Public Opinion Research [AAPOR], 2003; Winker, Menold, & Porst, 2013; Koczela at al., 2015). Recent developments in the use of paradata, the data about the process of generating survey data (Kreuter, 2013), provide tools for identifying the sources of non-unique records.

Some readers may be curious as to why the non-unique records reported in this paper had not been detected earlier by the 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 are simply insufficient. In recent research, finding duplicates was limited to small subsets of questionnaire items (Blasius & Thiessen, 2012) or establishing the likelihood of datasets containing duplicates (Kuriakose & Robbins, 2015).

Compared to Blasius & Thiessen (2012), our approach is *conservative* since we search for identical records over *all* answers and sometimes large numbers of variable values, as in the case of occupational codes, party preferences, and income brackets. Therefore, applying our procedure, we find a smaller number of duplicates but treating them as erroneous is even more justified. At the same time, we suggest that non-unique records be flagged (by a dummy variable) but not removed from the original data files. This suggestion is motivated by the need for preserving original data in order to assess the effect of non-unique records. It bears directly on the field of survey methodology.

We recommend that substantive analyses take into account duplications 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.

References

American Association for Public Opinion Research [AAPOR] (2003). Interviewer Falsification in Survey Research: Current Best Methods for Prevention, Detection and Repair of Its Effects. Retrieved October 17, 2015 from http://www.amstat.org/sections/srms/falsification.pdf

Biemer, P. 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*(July), 479–493.

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, p. 33). John Wiley & Sons.

Gideon, L. (Ed.). (2012). *Handbook of Survey Methodology for the Social Sciences*. New York: Springer.

Guidelines for Best Practice in Cross-Cultural Surveys (2010). Survey Research Center, Institute for Social Research, University of Michigan. Retrieved October 17, 2015 from http://www.ccsg.isr.umich.edu

Harkness, J. A., van de Vijver, F. J. R., & Mohler, P. Ph. (Eds.). (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*, 297–333.

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

Koczela, S., Furlong, C., McCarthy, J., & Mushtaq, A. (2015). *Statistical Journal of the IAOS*, *31*(3), 413–422.

Kreuter, F. (Ed.). (2013). *Improving Surveys with Paradata*. New Jersey: John Wiley & Sons.

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). Falsification in Surveys: Detecting Near Duplicate Observations. Retrieved October 17, 2015 from http://ssrn.com/abstract=2580502

Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., et al. (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. (forthcoming). Estimation bias due to duplicated observations.

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.

Smith, T. W. (2015). Resources for Conducting Cross-National Survey Research. *Public Opinion Quarterly, 79*(S1), 404-409.

Tomescu-Dubrow, I., & Slomczynski, K. M. (2014). Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Level Modeling in Cross-National Perspective. *ASK. Research & Methods, 23*(1), 103–114.

Treiman, D. J. (2009). *Quantitative Data Analysis. Doing Social Research to Test Ideas* (pp. 94-96). John Wiley & Sons.

Winker, P., Menold, N., & Porst, R. (Eds.). (2013). *Interviewers' Deviations in Surveys – Impact, Reasons, Detection and Prevention*. New York: Peter Lang , PL Academic Research.

Acknowledgments: This work was supported by a 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 Quantitative 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), and Jörg Blasius (Universität Bonn) for useful comments and remarks.

Table 1. Characteristics of 22 International Survey Projects.

Survey project[*]	Number of surveys	Number of countries[#]	Average number of questions	Average sample size	Number of cases
ABS	30	13	174	1456	43691
AFB	66	20	210	1499	98942
AMB	92	24	178	1645	151341
ARB	16	11	219	1230	19684
ASES	18	18	193	1014	18253
CB	12	3	275	2052	24621
CDCEE	27	16	299	1071	28926
CNEP†	8	8	294	1672	13372
EB†	152	37	342	913	138753
EQLS	93	35	167	1135	105527
ESS	146	32	223	1928	281496
EVS	128	50	347	1301	166502
ISJP	21	14	205	1229	25805
ISSP†	363	53	88	1359	493243
LB	260	19	251	1134	294965
LITS	64	35	636	1060	67866
NBB	18	3	172	1200	21601
PA2	3	3	271	1352	4057
PA8NS	8	8	345	1574	12588
PPE7N	7	7	299	2360	16522
VPCPCE	5	5	193	945	4723
WVS	184	89	221	1394	256582

	Number of all surveys	Number of distinct countries in all surveys	Average number of questions in all surveys	Average sample size in all surveys	Number of cases in all surveys
All surveys	1721	142	228	1330	2289060

[*] Data were downloaded at the turn of 2013/2014. For detailed dates and links to data sources, see https://dataverse.harvard.edu/dataverse/duprecords.

[†] For Comparative National Elections Project, Eurobarometer, and International Social Survey Programme, only selected survey editions were used.

[#] Countries or territories.

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).

Table 2. 17 International Survey Projects with Non-Unique Records.

Survey project	Number of non-unique records	Number of surveys	Number of countries	Average number of questions	Average sample size	Number of cases
				in affected surveys		
ABS	12	3	3	187	2430	7289
AFB	28	4	4	121	2273	9092
AMB	48	12	10	184	1869	22431
ASES	8	1	1	197	1000	1000
CB	2	1	1	261	1975	1975
CDCEE	168	3	3	296	1247	3740
EB	797	11	8	271	979	10773
EQLS	40	8	7	144	1069	8549
ESS	14	5	5	216	2045	10227
EVS	570	5	5	353	2045	10224
ISJP	2	1	1	235	1001	1001
ISSP	923	31	19	87	1922	59587
LB	1225	32	13	241	1114	35633
LITS	32	7	7	707	1000	7001
NBB	2	1	1	272	1987	1987
PPE7N	52	1	1	375	1769	1769
WVS	1970	36	31	227	1512	54449
	Number of non-unique records in all surveys	Number of all surveys	Number of all distinct countries	Average number of questions in all surveys	Average sample size in all surveys	Number of all cases
All surveys	5893	162	80	222	1523	246727

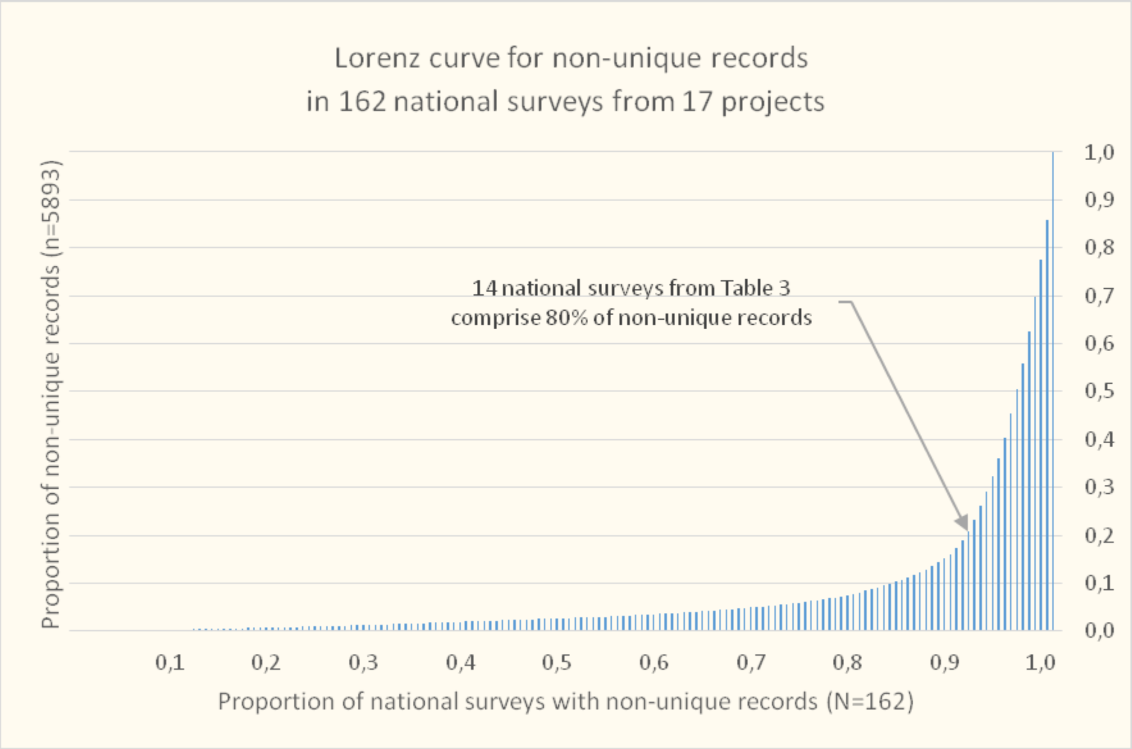
Lorenz curve for non-unique records
in 162 national surveys from 17 projects
Proportion of non-unique records (n=5893)
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
14 national surveys from Table 3
comprise 80% of non-unique records
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1,0
Proportion of national surveys with non-unique records (N=162)

Table 3. 14 National Surveys with the Largest Number of Non-unique Records

Project/wave	Country	Number of cases	Number of non-unique records	Proportion of non-unique records
CDCEE 1	Romania	1234	154	0.125
EB 19	Belgium	1038	148	0.143
EB 21	Belgium	1018	344	0.338
EB 31	Belgium	1002	220	0.219
EVS 1	United States	2325	528	0.227
ISSP 1989	Austria	1997	374	0.174
ISSP 1998	Bulgaria	1102	133	0.102
ISSP 2009	Norway	1456	160	0.110
LB 1996	Panama	1005	316	0.314
LB 2000	Ecuador	1200	733	0.611
WVS 1	Japan	1204	195	0.162
WVS 3	Mexico	2364	537	0.227
WVS 5	Ethiopia	1500	539	0.359
WVS 5	South Korea	1200	354	0.295

Table 4. An example of the regression analysis for the sample with included and excluded non-unique records: Interest in politics, dependent on gender, age, education, and respondent's reaction to the interview, South Korea 2005 (WVS 5)

Independent variables	**B**	**SE**	**Beta**	**P ≤**
Sample with included non-unique records, N = 1,200 Dependent variable: Interest in politics (scale recoded, 1 – low, 4 – high)				
Gender (1 = males, 2 – females)	-0.204	0.044	-0.133	0.000
Age (years)	0.008	0.002	0.141	0.000
Education (scale, 1 – lowest, 8 – highest)	0.030	0.015	0.073	0.041
Respondent's interest in interview (scale recoded, 1 – low, 3 – high)	0.125	0.033	0.109	0.000
Sample with excluded non-unique records, N = 846 Dependent variable: Interest in politics (scale recoded, 1 – low, 4 – high)				
Gender (1 = males, 2 – females)	-0.183	0.053	-0.119	0.001
Age (years)	0.006	0.002	0.112	0.010
Education (scale, 1 – lowest, 8 – highest)	0.017	0.018	0040	0.353
Respondent's interest in interview (scale recoded, 1 – low, 3 – high)	0.070	0.040	0.061	0.080

Appendix A. Homepages of the Twenty-Two International Survey Projects[*]

Project	Official name of project	Homepage
AFB	Afrobarometer	http://afrobarometer.org
AMB	Americas Barometer	http://www.vanderbilt.edu/lapop
ARB	Arab Barometer	http://www.arabbarometer.org
ABS	Asian Barometer	http://www.asianbarometer.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

[*] For some projects that do not have their own web pages, the archiving organization web page was used as a source.

Appendix B. Estimated Number of Publications Using Data from International Survey Projects

Project	Number of publications listed in			Number of Citations in Web of Science†
	homepages*	Google Scholar#	Web of Science†	
AFB[1]	428	1307 (5230)	55	204
AMB[2]	312	251 (502)	13	27
ARB[3]	30	174 (348)	3	6
ABS[4]	322	177 (354)	4	2
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**

* Data gathered on 2015-02-06.

# 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.

† Data gathered on 2015-03-31

The following expressions have been used for searches: [1] "afrobarometer" OR "afro-barometer" OR "afro barometer" [2] "americas barometer" [3] "arab barometer" [4] "asian barometer survey" [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"