Survey Research Methods (2023)
Vol. 17, No. 2, pp. 131-145
doi:10.18148/srm/2023.v17i2.7961
European Survey Research Association

Detecting Fabricated Interviews Using the Hamming Distance

Jörg Blasius and Lukas Sausen
University of Bonn
Institute for Political Science and Sociology

In the research literature on survey methodology, there is considerable discussion of interviewer effects and how to prevent data fabrication; however, there is little discussion of the detection of data fabrication by interviewers in published data, and there are even fewer papers examining the phenomenon of employees of survey research organizations fabricating data. Among them, Blasius and Thiessen (2015) show for the PISA 2009 principals' data that employees of survey research organizations in some countries duplicate cases to generate data. While the authors focus there on exact copies, more sophisticated data fabrication techniques might include duplicating whole cases and subsequently changing a few entries. By calculating Hamming distances and applying them to the same data, we show that—in some countries in particular—large parts of the data have been duplicated, and most of them have been retrospectively modified to a small degree.

Keywords: Fabricated data; string distances; PISA data

1 Introduction

In the literature there is a great deal of discussion of interviewer effects and respondent behaviour. Many papers concentrate on interviewees whose response behaviour consists of giving merely sub-optimal answers, a phenomenon that is known as satisficing (Krosnick, 1991; Krosnick & Alwin, 1987), or strong satisficing (Krosnick et al., 1996) in instances where interviewees give arbitrary answers. Another widespread topic of discussion concerns interviewer effects, and specifically how to prevent interviewers from fabricating data. The discussion began with Crespi (1945, p. 431) claim that the "cheater problem is essentially a moral one", whereby cheating "lies as much in the structure of the ballots and the conditions of administration as in the personal integrity of the interviewer". He discusses several questionnaire problems that might demoralize interviewers, e.g., unreasonable length, too many "why's and what for's" (Crespi, 1945, p. 438), apparent repetition of questions, lengthy wording, or complex, difficult, and antagonizing questions. Almost 60 years later, Groves (2004, p. 2) provides a broad definition of interviewer falsification, which "means the intentional departure from the designed interviewer guidelines or instructions, unreported by the interviewer, which could result in the contamination of data. 'Intentional' means that the interviewer is aware that the action deviates from the guidelines and instructions." They discuss various control strategies and state that the "various control practices are actively followed in most survey organizations, so the prevalence of falsification is quite low" (Groves, 2004, p. 2). For a recent examination of interviewer effects and how to monitor interviewers see, among many others, the reader by Olson et al. (2020).

There is, in contrast, very little discussion of the detection of data fabrication by interviewers and other survey organization employees; scholars and researchers seem to believe that (1) survey institutes take care that their control mechanisms are sufficient for detecting fabricated interviews, and (2) employees of survey research organizations do not fabricate data. Both assumptions do not hold: Data fabrication by interviewers and by other survey organization employees does occur (Blasius, 2018; Blasius & Thiessen, 2012, 2015, 2021; Cohen & Warner, 2021; Hernandez et al., 2022; Koczela et al., 2015; Kuriakose & Robbins, 2016; Schafer et al., 2005; Slomczynski et al., 2017; Yamamoto & Lennon, 2018) and sometimes occurs at levels that must necessarily undermine confidence in survey results. It is therefore important that principal investigators and survey managers have the tools to detect data fabrication. This paper discusses a new method for detecting fabricated data—the Hamming distance (HD)[1]—and we demonstrate its advantages over previous approaches.

According to Blasius and Thiessen (2012), there are three actors involved in the data collection process: interviewees,

Contact information: Jörg Blasius, University of Bonn, Institute for Political Science and Sociology, Lennéstr. 27, 53113 Bonn, Germany (E-mail: jblasius@uni-bonn.de)

[1] To the best of our knowledge, the idea to use the Hamming distance for detecting duplicates in survey data was first aired by Powałko (2015).

interviewers, and other employees of survey research organizations. In all of these three groups there are individuals whose intention may be to minimize their own time and energy commitment. While interviewees who are (strong) satisfiers prefer to invest as little time as possible in the interview (Krosnick, 1991), interviewers might actually be acting rationally in conducting as many interviews as possible as quickly as they can (Blasius & Thiessen, 2021). The same should be true for other employees of survey research organizations, obligated as they are to collect a certain number of interviews in a certain time; the person(s) responsible for the field work must deliver the contracted number of interviews to a fixed deadline—and may duplicate data to do so. To provide a general framework encompassing potential wrongdoing on the part of all three actors who might impact data quality (strong satisfiers, interviewers who (partly) fabricate their interviews, and survey organization employees who generate data via copy-and-paste), Blasius and Thiessen (2012) propose the term "simplification".

Research organization employees who have access to the electronic data file can simplify their assignments by fabricating entire interviews, in the simplest case via copy-and-paste (Blasius, 2018; Blasius & Thiessen, 2015). For PISA (Program for International Student Assessment), Blasius and Thiessen (2015) and Blasius (2018) show that, in certain countries, large parts of the school principals' questionnaires had been duplicated, sometimes even triplicated and quadrupled. The authors found the largest share of such duplications in Slovenia, the United Arab Emirates (UAE), and Italy.

Since school principals' data in PISA are obtained via self-administered questionnaires, only two sources of simplification are possible: First, the respondents may simplify their tasks by utilizing simple response patterns such as straight-lining or giving stereotypical answers; second, the employees of survey research organizations may simplify their tasks by duplicating cases.

In this paper, we use the principals' data from PISA 2009 (OECD, 2009) and discuss a new methodological approach to explore the process and scope of data fabrication via copy-and-paste. While Blasius and Thiessen (2012, 2015), Slomczynski et al. (2017) and Blasius (2018) consider only complete copies, i.e., the simplest form of data fabrication, we also include copies in which minor changes were made, i.e., by changing a few values in a large set of items. Using the Hamming distance (Hamming, 1950), we will present an approach for detecting not only complete duplicates, but also those where single entries have been modified. The entire distribution of all variables from all respondents is taken into account, thus enabling the construction of statistical tests to detect whether or not the similarities between two interviews can be explained by random chance.

The dataset is publicly available, well known, and widely used within the social and educational sciences. Furthermore, in many of the participating countries, the PISA surveys' results, which are collected under the aegis of the OECD, make headlines in the national newspapers and TV news.

2 State of Research

Among the recent approaches for detecting fabricated interviews are scaling techniques, especially principal component analysis (PCA) and multiple correspondence analysis (MCA)—the latter method is also known as homogeneity analysis (Gifi, 1990). Both methods provide factor loadings, factor scores, eigenvalues, and explained variances. When applying these techniques to the detection of fabricated interviews, the focus is not on a specific, substantive solution, but solely on the factor scores. If they are the same on all dimensions the cases are identical.

2.1 Detecting identical response patterns

Blasius and Thiessen (2012, p. 64) apply MCA to a set of 36 variables in successive order from the World Value Survey 2005–2008 (v60–v95, in the public use file), including "eight four-point variables on gender roles, one 10-point satisfaction with financial situation variable, two ranking scales of four choices on national goals, six ranking scales of four choices on materialism and post-materialism values, 10 six-point self-description variables, and four 10-point variables on technology". It is evident that most of these items are not inter-correlated. But regardless of the questions' content, the same responses to all 36 items provide the same factor scores on all dimensions. A visualization of the frequencies of each nation's factor scores with bar charts shows that, as expected, most of the countries did not exhibit any identical response patterns (IRPs) such as duplicates, triplets, or quadruplets. However, some countries showed many IRPs, among them South Korea, Ethiopia, and India (Blasius & Thiessen, 2012, p. 66).

In a comprehensive study, Slomczynski et al. (2017) analysed 1,721 national surveys in 22 international projects, covering 142 countries and 2.3 million respondents. They found a total of 5,893 duplicates (which they call non-unique responses or NURs) concentrated in 162 national surveys, in 17 projects and 80 countries. Further, 80% of all NURs are present in just 14 surveys, while the remaining 148 surveys contain 20% of the NURs (Slomczynski et al., 2017, p. 5).

In social surveys in which both interviewers and other employees of survey research organizations are involved in gathering the data, both actors can be responsible for IRPs. In PISA, besides the student assessments, the principals (or their substitutes) of the randomly selected schools were asked to fill in a self-administered questionnaire concerning the conditions at their school. This information is needed for all analyses that involve schools' characteristics, for example school equipment and teachers' abilities. Because there are no interviewers, IRPs in the respective datasets can be only manufactured by those employees of survey research organizations who have access to the electronic data file, which might be a single individual.

2.2 Data fabrication by survey organization employees

Distributed throughout a total of 18,233 cases, across 184 items—making $1.729 \cdot 10^{72}$ theoretically possible response patterns—Blasius and Thiessen (2015) found 101 different IRPs (91 duplicates, 8 triplets, 2 quadruplets, in total 214 cases) in the 2009 PISA principals' data. Since the probability of even the one-off re-occurrence of a pattern is already vanishingly small ($5.78 \cdot 10^{-71}$), the observed quantity of IRPs is simply so improbable for a total of 18,233 cases that any explanation other than systematic data manipulation is unimaginable (Blasius & Thiessen, 2015, p. 486). The uneven distribution of IRPs among countries further supports the assumption of deliberate data fabrication via copy-and-paste. No single IRP arose *across* the 71 countries that participated in the study, and the great share of IRPs occurred *within* the three following countries: Italy, Slovenia, and the UAE. Taking the example of Slovenia, Blasius and Thiessen (2015, p. 487) observe that approximately 18% of all Slovenian cases included IRPs, rendering at least half of those observations useless as mere duplicates.

In theory, a few duplicates might be "legitimate", for example, when there are two schools in the same building, as reported in the Technical Report (OECD, 2012) for Slovenia; instead of requesting the answers from both principals, one questionnaire might be copied-and-pasted to save time and conceal the original error. But this assumption does not hold: When reducing the item set to 40 variables, Blasius and Thiessen (2015, pp. 487, 489) find in total 146 duplicates, 19 triplets, and 4 quadruplets—again, the large majority of this increase in IRPs can be ascribed to Slovenia, the UAE, and Italy. For example, in Slovenia the number of duplicates increased from 20 to 41, the number of triplets from 6 to 14, and the number of quadruplets from 1 to 2. This is another very strong indicator of data fabrication by employees of survey research organizations. Further, the strong increase of IRPs in Slovenia, the UAE, and Italy is an indicator that some variables were modified after entire cases were duplicated.

Unlike academic researchers, surveying institutes may have an additional incentive for cheating through data duplication. Kuriakose and Robbins (2016, pp. 283–284) argue that if "a dishonest firm carries out a sufficient number of interviews among a diverse—and reasonably representative—segment of the target population, then the results will generally yield both the expected distributions on known variables and the proper correlations between variables. If the observations in this partial survey are duplicated one or more times, then the required survey sample size is reached at a substantially lower cost."

Like other companies, survey research organizations should not be conceived of solely as uniform agents, but also as groups of individual actors. When forms of task simplification are observed on an institutional level, this does not necessarily imply that data fabrication is deliberately imposed by management. Several constraining factors can prompt individual employees into illicit task simplification as well: looming deadlines, unfeasible response rate requirements, underperforming subordinated interviewers, inconclusive data, etc. A single employee—or indeed a team or an entire department—might "take matters into their own hands". Without a close assessment of the specific situational context, it is not possible to tell where and when within survey research institutes intentional data fabrication occurs.

2.3 The Kuriakose and Robbins approach: Percentmatch

Kuriakose and Robbins (2016) conducted a meta-analysis across multiple studies with a total of 1,008 country-year-surveys as a basis. Within these datasets, they analysed the proportions of nearly identical response patterns (NIRPs) via the Stata module Percentmatch (Kuriakose, 2015). This module can be used to calculate the highest percentage match (near duplicates) between observations.

Percentmatch provides a similarity index by dividing the number of identical items from two observations by the number of total assessed variables and multiplying this value by 100, resulting in a percentage between 100 and 0 for the degree of matching. To give an example, assuming two cases and 100 variables, imagine 95 of them have the same responses in both cases; the remaining five are different, resulting in a similarity index of 95%. This procedure is reiterated for every possible combination of observations, and ultimately the highest percentage match value for each case is saved in the dataset (Kuriakose & Robbins, 2016, p. 284), i.e., the resulting number of values is equal to the number of cases in the dataset.

The authors use 85% as threshold for potential falsification. From 1,008 surveys examined, they find that 35.8 percent of the datasets contained no violation of the 85% criterion at all, 46.8% of the surveys included up to 5% NIRPs, 7.2% of surveys included between 5 and 10% NIRPs, and 10.1% of surveys consisted of more than 10% potential falsifications (Kuriakose & Robbins, 2016, p. 287). Simmons et al. (2016) demonstrated that the method proposed by Kuriakose and Robins overestimates the number of fabrications, since the proposed match statistics are "extremely sensitive to the number of questions, number of response options, number of respondents, and homogeneity with the population" (Simmons et al., 2016, p. 327). Furthermore, Percentmatch does not provide users with the possibility to determine whether a NIRP can be flagged because of an outlying similarity between two cases.

In the following, we use a program that implemented the so-called Hamming distance (HD) to address the aforementioned shortcomings of the research design decisions by Kuriakose and Robbins (2016). While the HD is conceptually the same measure underlying Kuriakose and Robbins's Percentmatch-approach, the actual implementation allows us to evaluate all pairwise similarities of the observations of the dataset, and to consider the full distribution of the similarities without using a predefined threshold. The more variables the dataset contains, the more categories the variables have, and the more homogeneous they are, the more the cases can differ. In addition to the number of variables and variable categories, as well as the sample size, the homogeneity of the sample has an impact on the distribution of similarities. The distribution of similarities can then be used to check whether individual similarities can be considered outliers. It is not necessary to determine either a (uniform) threshold or the number of cases (that is, the number of clear similarities between two cases) which lie above the chosen threshold. The statement that two cases are "too similar" can be made on the basis of the estimated probability of the occurrence of this similarity.

3 Data

For the empirical section of this paper, we use the principals' data from PISA 2009, also used by Blasius and Thiessen (2015). The data were collected to gather information on the level of individual schools—for example, on resource shortages, management practices, and computer facilities. The newly downloaded dataset contains 18,641 cases from 72 countries (OECD (2009); for details such as sampling procedures, see OECD (2012).

Blasius and Thiessen (2015) and Blasius (2018) concentrate on IRPs that were detected by way of identical factor scores. However, employees of survey research organizations might duplicate cases and change a few entries afterwards to hide their aberrant behaviour. Cases that have been duplicated and afterwards slightly modified can be detected by reducing the set of variables and searching for very similar factor scores on a large number of dimensions. As an example, we present for the Slovenian PISA data a subset of 34 variables (SC11_1 to SC14_5, variable names as given in the public use file) and 24 cases (Table 1, the columns contain the variables, the rows contain the interviews, the cells contain the numerical responses to the questions—differences between the cases are marked).

Table 1 shows several IRPs and some NIRPS with one, two, and three changes between the cases. For example, ID-271 differs from ID-16 in the items SC13_12 and SC13_13 (the respective values are interchanged), and ID-247 and ID-48 are different in the items SC14_1 and SC14_5. The method applied in Table 1—of searching for similar factor scores in the first two, three, or four dimensions—can be used to illustrate the problem in a small subset of the data, but it is not possible to detect a majority of NIRPs in a large dataset when a large share of variables is to be included in the analysis.

The following analyses are based on 202 items; included are the questions SC01Q01 to SC02Q01, SC04Q01, SC05Q01, and SC11Q01 to SC27Q01 (variable names as given in the public use file). To restrict the dataset to variables with fewer than 11 categories (with values running from 0 to 9), we exclude the few variables with metric information. The set of selected variables consists of 139 dichotomous questions, 17 questions with three response categories, 40 questions with four response categories, and six questions with five response categories, resulting in $2^{139} \cdot 3^{17} \cdot 4^{40} \cdot 5^{6} = 1.70 \cdot 10^{78}$ possible response patterns. Including item nonresponse, the total number of possible combinations rises to $3^{139} \cdot 4^{17} \cdot 5^{40} \cdot 6^{6} = 1.52 \cdot 10^{109}$, an almost inconceivably large number. Using the HD, we illustrate how the strings thus generated can be used to identify IRPs and NIRPs.

4 Method

In order to detect IRPs and NIRPs, the distances between all cases have to be examined. The central concept of our approach is to treat the total set of selected variables as a string of digits, where each variable is represented by a single alphanumeric character within the string. For each case in the dataset, we create a string which will be compared against all other strings, character by character. Thereby, one single character represents one single manifestation of a variable in a sequence of digits (= alphanumeric characters).

All variable values, i.e., all "characters", of a case are merged into one single string. The transformation of the aforementioned 202 items in the PISA 2009 principals' dataset results in strings with a length of 202 characters. In the next step, we compare each string with every other string to compute the distance between them. The more characters that are identical at the same position, the greater the similarity between the strings.

A commonly used technique to acquire the information described above is a string metric, which provides a numeric similarity index to show the relationship between two strings. String distances offer a way of quantifying how (dis)similar sequences of symbols (numbers, letters, ...) are related to one another by counting the minimum number of operations that is required to transform one string into the other (Navarro, 2001; Van der Loo et al., 2014).

In the following we explain how the HD algorithm operates. An HD requires strings of equal length for a comparison of characters, position by position. The number of positions where symbols do not match each other defines the distance between two strings. Consequently, the measure specifies the minimum number of substitution operations re-

Table 1

Excerpt from the data matrix, principal data, PISA 2009, Slovenia

	SC11													SC12		SC13														SC14				
ID	1	2	3	4	5	6	7	8	9	10	11	12	13	1	2	1	2	3	4	5	6	7	8[b]	9	10	11	12	13	14	1	2	3	4	5
140	3	2	2	4	1	1	1	2	1	1	1	2	1	3	2	1	2[a]	1	1	2	1	1	7	1	1	1	1	1	1	2	2	2	1	1
83	3	2	2	4	1	1	1	2	1	1	1	2	1	3	2	1	1[a]	1	1	2	1	1	7	1	1	1	1	1	1	2	2	2	1	1
292	3	2	2	4	1	1	1	2	1	1	1	2	1	3	2	1	1[a]	1	1	2	1	1	7	1	1	1	1	1	1	2	2	2	1	1
200	1	1	1	1	1	2	1	1	1	1	1	1	2	2	3	1	1	1	1	1	2	1	7	1	1	1	1	1	1	1	2	2	2	2
212	1	1	1	1	1	2	1	1	1	1	1	1	2	2	3	1	1	1	1	1	2	1	7	1	1	1	1	1	1	1	2	2	2	2
325	1	1	1	1	1	2	1	1	1	1	1	1	2	2	3	1	1	1	1	1	2	1	7	1	1	1	1	1	1	1	2	2	2	2
271	1	1	1	1	1	1	1	1	1	1	1	1	1	3	2	2	1	1	1	2	2	1	7	1	1	1	2[a]	1[a]	2	1	2	2	2	2
16	1	1	1	1	1	1	1	1	1	1	1	1	1	3	2	2	1	1	1	2	2	1	7	1	1	1	1[a]	2[a]	2	1	2	2	2	2
41	1	1	1	1	1	1	2	3	3	1	2	1	2	3	2	2	2	1	2	2	2[a]	1	7	1	1[a]	2[a]	1	1	2	1	1	2	2	2
18	1	1	1	1	1	1	2	3	3	1	2	1	2	3	2	2	2	1	2	2	7[a]	1	7	1	2[a]	1[a]	1	1	2	1	1	2	2	2
218	1	1	2	2	1	2	2	2	1	2	2	2	2	3	3	1	2	1	1	2	2	1	7	1	1	1	1	1	2	1	1	2	2	2
275	1	1	2	2	1	2	2	2	1	2	2	2	2	3	3	1	2	1	1	2	2	1	7	1	1	1	1	1	2	1	1	2	2	2
98	1	1	1	1	1	1	2	2	2	1	2	2	2	3	3	2	2	2	2	2	2	1	7	1	1	1	2	1	2	2	2	2	2	2
134	1	1	1	1	1	1	2	2	2	1	2	2	2	3	3	2	2	2	2	2	2	1	7	1	1	1	2	1	2	2	2	2	2	2
282	1	1	1	1	1	1	2	2	2	1	2	2	2	3	3	2	2	2	2	2	2	1	7	1	1	1	2	1	2	2	2	2	2	2
119	1	2	2	2	1	1	2	3	2	2	3	2	2	3	3	2	1	1	2	2	2	1	7	1	1	1	1	2	2	1	2	2	2	2
160	1	2	2	2	1	1	2	3	2	2	3	2	2	3	3	2	1	1	2	2	2	1	7	1	1	1	1	2	2	1	2	2	2	2
273	1	2	2	2	1	1	2	3	2	2	3	2	2	3	3	2	1	1	2	2	2	1	7	1	1	1	1	2	2	1	2	2	2	2
54	1	1	1	1	1	1	1	2	1	1	1	1	1	3	3	2	2	1	2	2	2	1	7	1	1	1	1	1	2	1	2	2	2	2
319	1	1	1	1	1	1	1	1	1	1	1	1	1	3	3	2	2	1	1	1	1	1	7	1	1	1	1	1	2	1	1	2	2	2
247	1	1	1	1	2	1	2	2	1	1	2	2	2	2	1	1	2	1	2	2	1	1	7	1	1	1	1	1	2	2[a]	2	2	2	2[a]
48	1	1	1	1	2	1	2	2	1	1	2	2	2	2	1	1	2	1	2	2	1	1	7	1	1	1	1	1	2	1[a]	2	2	2	1[a]
198	1	1	1	3	1	1	2	3	2	1	2	2	2	3	3	1	2	1	1	1	1	1	7	1	1	1	1	2	2	1	1[a]	2	2	2
46	1	1	1	3	1	1	2	3	2	1	2	2	2	3	3	1	2	1	1	1	1	1	7	1	1	1	1	2	2	1	2[a]	2	2	2

Reading example: Raw data for SC11 refers to the item battery Q11: "Is your school's capacity to provide instruction hindered by any of the following issues?" First column then refers to the first item of the battery: "A lack of qualified science teachers", the four response categories are: 1 = "not at all", 2 = "very little", 3 = "to some extent", 4 = "a lot"; see "School Questionnaire for PISA 2009, Main Survey, December 2008", page 10). The principal with school ID 140 (first row) marked "3 = to some extent".

[a] Responses that are different within a set of copied and pasted data. [b] Item 8 of question SC13 was not asked for in Slovenia and therefore coded as missing (7).

quired to transform string A into string B. Table 2 shows some examples to demonstrate the concept. For example, the strings "distaste" and "distance" are identical except for two letters—the "s" in position 6 has been replaced by "n", and in position 7 the "t" has been replaced by "c"; between these two strings, HD = 2. The words "rocks" and "flock", while sharing three common letters, have no character matches in the same position, so HD = 5. The last example uses concatenated digits; between "246754" and "247549", HD = 4. Of the six characters, both strings share the first two and differ in the last four positions.

Applying the method to survey data, a dataset may consist of several thousand strings (cases) with one hundred and more characters (variables in the questionnaire). Since all strings are of the same length and each position contains the same variable among all cases, a pairwise comparison of characteristics is possible. Thereby, missing values can be included in the same way as substantive information, namely as numbers in a string of characters. For example, the characters "1" to "5" symbolize the valid responses in a five-point scale, "8" and "9" reflect different possibilities of response refusals, and "0" symbolizes "not applicable".

Because this distance measure a) provides information regarding similarity based on characters and b) the strings generated here contain one character per survey item, we are able to assess the similarity of records by the number of identical characters. Thereby, two strings with HD = 0 comprise an IRP that could also be detected via PCA or MCA. However, the HD not only provides us with the ability to identify IRPs; it also indicates differences in one or more characters (variables) between two cases. In general, HDs can serve as a measure of similarity with a ratio scale for survey data; their possible values range from zero to K, where K is the number of variables, i.e., the length of the string.

HDs can be calculated for each pair of cases in the dataset; the number of pair comparisons is equal to $\frac{N(N-1)}{2}$, with N being the number of cases in the dataset. If N is large enough, if the cases have been collected independently (which is assumed in most survey data), and if the variables are uncorrelated to each other (the great majority of variables within a dataset is uncorrelated), the resulting HDs follow the normal distribution, with a mean value of $\overline{x}$, i.e. the average number of identical characters between two strings, and a standard deviation of s_x. This property enables the detection of outlying similarities within the distribution of HDs (e.g., the detection of highly similar cases—IRPs and NIRPs), depending on the mean, the standard deviation, and the number of pair-comparisons. For the following calculations we use the R package stringdist (Van der Loo et al., 2014), and to visualize the marginal distributions of the HDs we give bar charts and Q-Q plots; the r-code is given in the Online Appendix.

5 Empirical Results

Using the example of PISA 2009, we computed the HDs for the aforementioned set of 202 items for the entire dataset, as well as for each individual country. Figure 1 shows the frequencies of HDs as a bar chart for all cases in the PISA 2009 principals' survey, with less than 20% of missing values within the set of 202 items (we thus excluded 486 cases, or 2.7%). In total, 17,809 cases are included, resulting in $\frac{N\cdot(N-1)}{2} = 158,571,336$ observed distances. The bar chart depicts a normal-looking distribution with a mean of 97.4 HDs, a standard deviation of 11.9, a median of 97, a mode of 97 (consisting of 5,648,774 pair comparisons), and a range from 0 to 177.

Since the bar chart depicts almost 160 million pair-by-pair comparisons, the number of comparisons symbolized on the y-axis exceeds more than 5,500,000 values for HDs close to the mean (Figure 1). Despite being very short, still visible on the very left of Figure 1 are the bars symbolizing outlying HDs with values between 0 and 6 (note the first tick-mark on the y-axis symbolizes 1,000,000 cases). Using the 95% confidence interval of [74; 121], there are roughly eight million HDs outside the interval. For detecting outlying HDs, one has to take into account that the number of pair comparisons is extremely large. Assuming five standard deviations on both sides of the distribution, the probability of finding a value by coincidence on either the left or right side is $p = 5.73 \cdot 10^{-7}$. This is a very small value, but 91 HDs are still expected to be outside of the corresponding 99.999999% interval. Therefore, we increase the confidence interval to six standard deviations.

The probability of finding an outlying HD decreases to $p = 9.87 \cdot 10^{-10}$, the number of expected outlying HDs decreases to 0.3, and the respective confidence interval is [26; 169]. However, there are some pair comparisons where HD $\geq$ 169 (see Figure 1), and there is a relatively large number of cases with HD $\leq$ 26. Although the large majority of pair comparisons occur between the 72 countries (cross-country combinations), there is neither an IRP nor a NIRP with an HD $\leq$ 20 *between* the countries—all IRPs and NIRPs occur *within* the countries. Table 3 shows the frequency of HDs within selected countries.

As Table 3 shows, there are 354 occurrences of HD*s* $\leq$ 20; the great majority, namely 290 instances (82 percent), can be observed within the following three countries: Slovenia, the UAE, and Italy. When increasing the threshold to HD $\leq$ 30, the value for Italy increases from 51 to 102, for Slovenia from 175 to 253, and for the UAE from 64 to 113 (Online Appendix, Table A). The scarcity or nonappearance of IRPs and NIRPs in other countries, and the absence of HD $\leq$ 20 between countries, strongly suggest a non-random distribution of IRPs and NIRPs, one which cannot be a product of mere chance. These results correspond strongly with Blasius and Thiessen (2015) findings: The authors especially identi-

Table 2

Examples for the Hamming Distance HD

String A	distaste	solitarily	rocks	246754
String B	distance	similarity	flock	247549
Substitutes	s/n, t/c	o/I, l/m, t/l, l/t	r/f, o/l, c/o, k/c, s/k	6/7, 7/5, 5/4, 4/9
HD	2	4	5	4

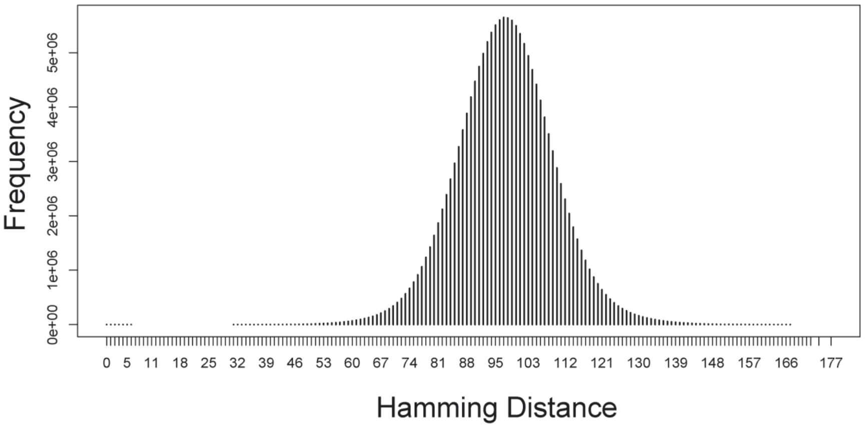


Figure 1

PISA 2009, Hamming distances between all principals, bar chart, $K = 202$ *variables.* $N = 17,809$ *cases with less than 20% missing values; Number of HDs* $= \frac{17,809 \cdot 17,808}{2} = 158,571,336$*; Mean* $= 97.4$*; Std.Dev.* $= 11.9$

fied Slovenia, the UAE, and Italy as countries in which parts of the data were duplicated.

In the next step we give the bar charts for the HDs in individual countries. As examples for legitimate, "unsuspicious" results, we use the results of Canada, Germany, and Luxembourg (Figure 2). Overall, the subsets of the countries have a normal-looking distribution of HDs: Canada's ($N = 963$) bar chart is almost smooth, the one for Germany ($N = 208$) differs slightly from the normal distribution, and while Luxembourg ($N = 39$) shows some fluctuation, this is due to the small number of cases. The means of these countries are smaller than those of the entire sample, which was expected, since school characteristics are more similar within countries than between countries.

Figure 3 gives the bar charts for the HDs of Italy, Slovenia, and the UAE. It can be seen that Italy and Slovenia exhibit a normal-looking distribution as well, with mean values and standard deviations close to the three legitimate countries mentioned above. In contrast, the bar chart for the UAE shows a "capped" maximum, with a higher mean and a higher standard deviation than is the case for all other countries. One explanation for these findings is that the survey research organisation employees created data records randomly.

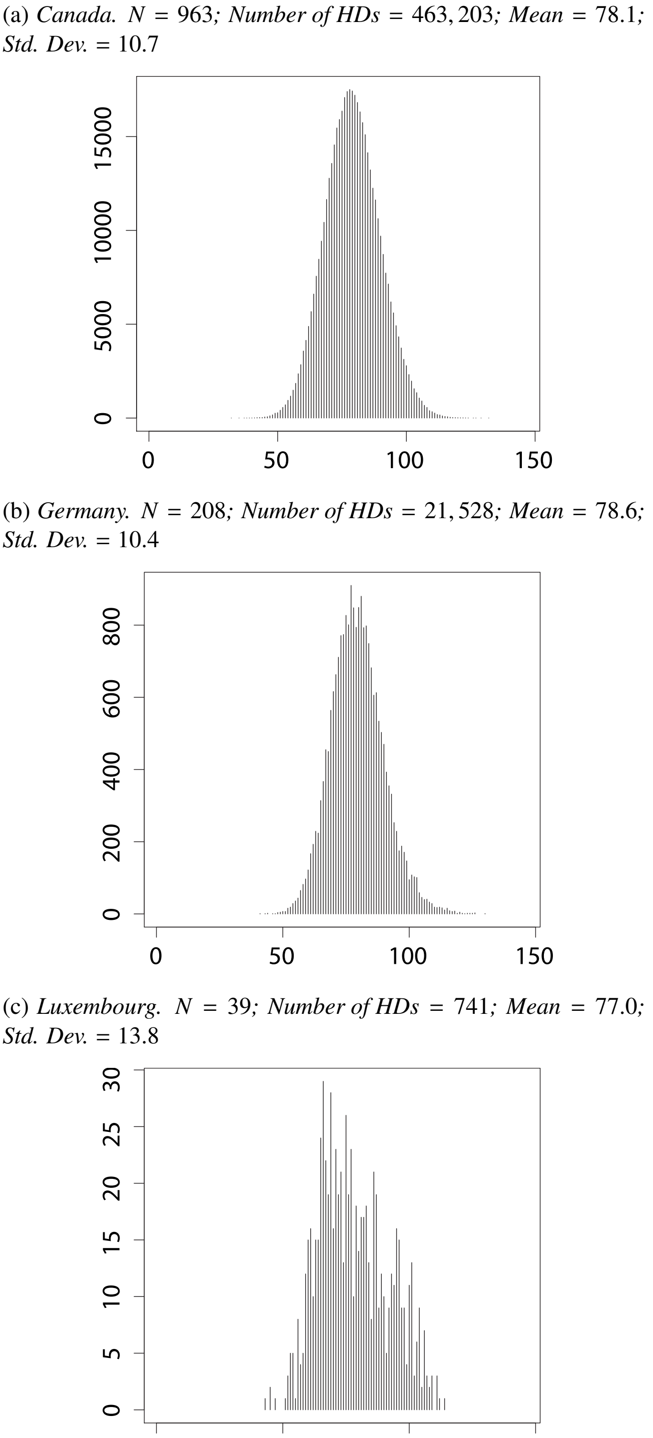
Furthermore, Slovenia and the UAE have a barely noticeable local maximum of HDs close to 0 that reflects the solutions already shown in Table 3. In the case of Italy, the course of the graph resembles a flat line towards 0: The 12 occurrences of IRPs and the 39 occurrences of NIRPs are hardly noticeable in the bar chart due to the relatively large sample size for this country ($N = 1037$, number of HDs $= 538,203$). The bar charts of the HDs and the numerical solutions (HDs ≤ 30) for all countries are shown in the Online Appendix (Figure A, Table A).

For a better graphical detection of outliers, we apply Q-Q plots plotting the theoretical quantiles against the sample quantiles. The diagonal line reflects the expected normal distribution of the HDs. Figure 4 shows the data for Canada, Germany, and Luxembourg. For Canada the data fits almost perfectly in terms of the expected normal distribution. For Germany there are a few outlying HDs in the upper parts of the HDs, which are scarcely visible in Figure 2: One possible explanation is that some schools are quite different in comparison to the rest of the national subsample, another explanation is that some principals took little care when answering the questions—they might be (strong) satisfiers. The deviations for Luxemburg are probably due to the small number of cases.

Table 3

Hamming distances in PISA 2009, shown are the frequencies for $HD = 0$ to $HD = 20$, selected countries

Distance	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Sum
Austria	2	0	2	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7
Belgium	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	4
Colombia	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
Hungary	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Italy	12	10	5	1	2	0	0	4	2	3	0	0	2	1	4	1	0	1	2	1	0	51
Latvia	1	0	2	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	5
Mexico	3	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	5
Netherlands	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	3
Qatar	2	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Slovenia	5	28	36	16	12	11	9	5	5	5	5	7	4	3	4	3	5	3	2	2	5	175
Spain	4	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7
Switzerland	2	0	0	0	1	0	2	0	1	0	0	0	0	0	1	0	0	0	0	0	0	7
TTO	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2
UAE	31	15	5	6	2	0	0	1	0	2	2	0	0	0	0	0	0	0	0	0	0	64
Uruguay	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	4
Other C.s	4	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	1	1	0	8
Total	73	55	53	25	18	14	15	11	10	12	10	8	7	4	9	4	6	6	5	4	5	354

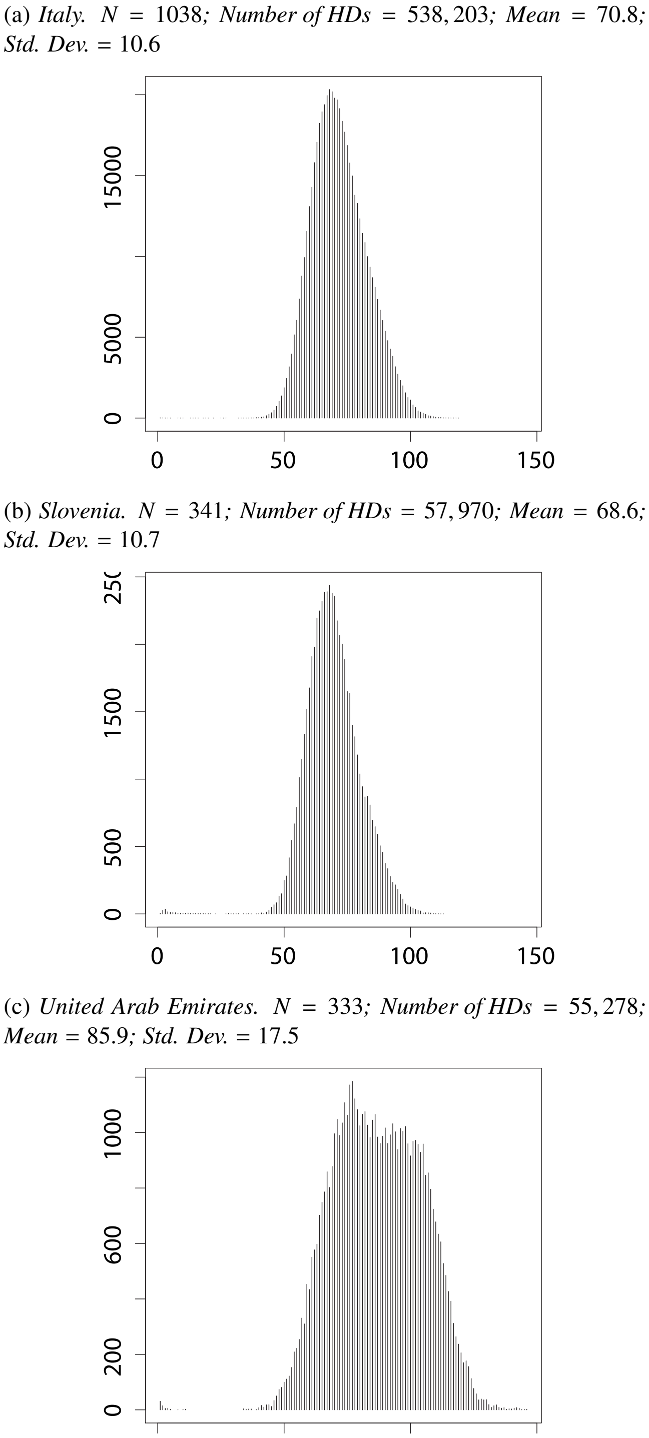

(a) *Canada. N* = 963*; Number of HDs* = 463, 203*; Mean* = 78.1*; Std. Dev.* = 10.7

(b) *Germany. N* = 208*; Number of HDs* = 21, 528*; Mean* = 78.6*; Std. Dev.* = 10.4

(c) *Luxembourg. N* = 39*; Number of HDs* = 741*; Mean* = 77.0*; Std. Dev.* = 13.8

Figure 2

Hamming distances, bar charts, selected unsuspicious countries

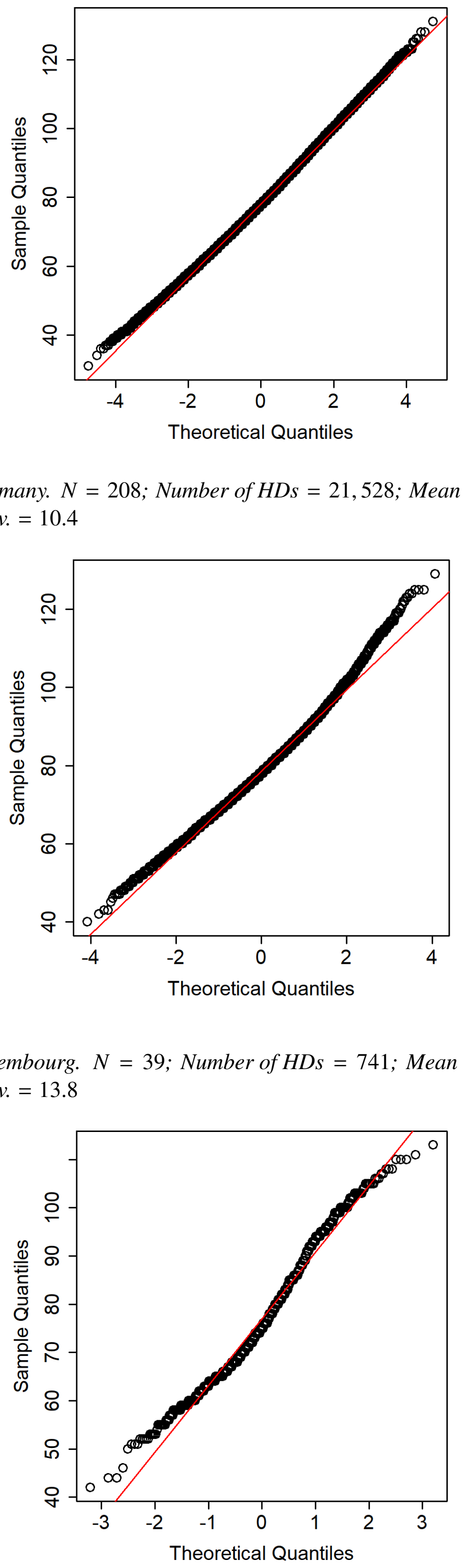

(a) *Italy. N* = 1038*; Number of HDs* = 538, 203*; Mean* = 70.8*; Std. Dev.* = 10.6

(b) *Slovenia. N* = 341*; Number of HDs* = 57, 970*; Mean* = 68.6*; Std. Dev.* = 10.7

(c) *United Arab Emirates. N* = 333*; Number of HDs* = 55, 278*; Mean* = 85.9*; Std. Dev.* = 17.5

Figure 3

Hamming distances, bar charts, selected suspicious countries



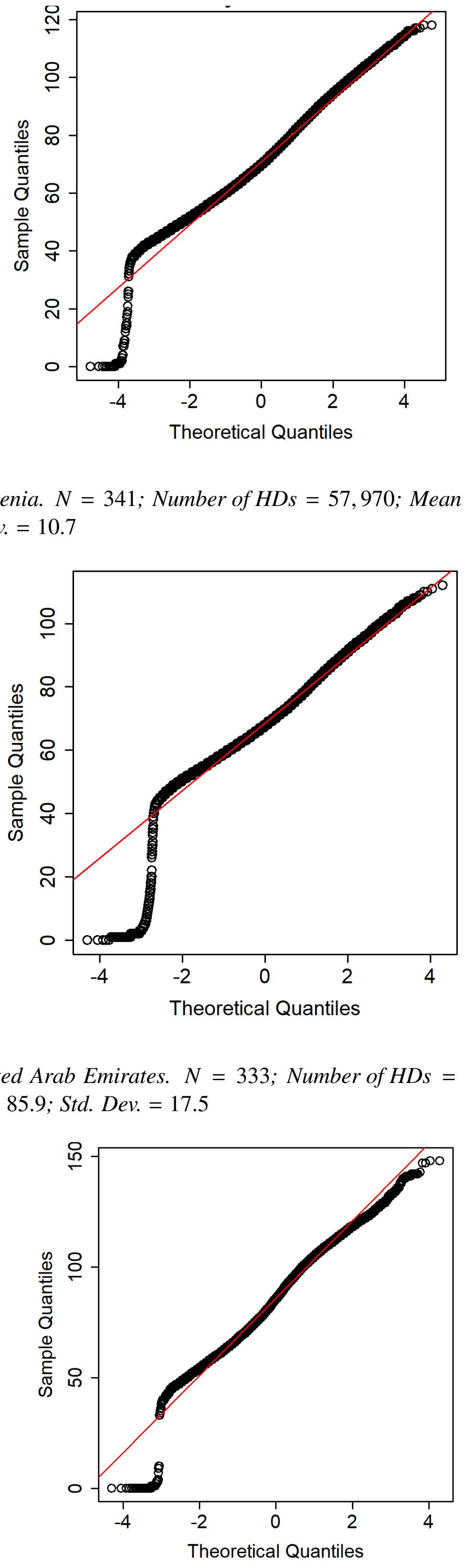
(a) *Canada. N* = 963; *Number of HDs* = 463, 203; *Mean* = 78.1; *Std. Dev.* = 10.7

(b) *Germany. N* = 208; *Number of HDs* = 21, 528; *Mean* = 78.6; *Std. Dev.* = 10.4

(c) *Luxembourg. N* = 39; *Number of HDs* = 741; *Mean* = 77.0; *Std. Dev.* = 13.8

Figure 4

Hamming distances, Q-Q-plots, selected unsuspicious countries



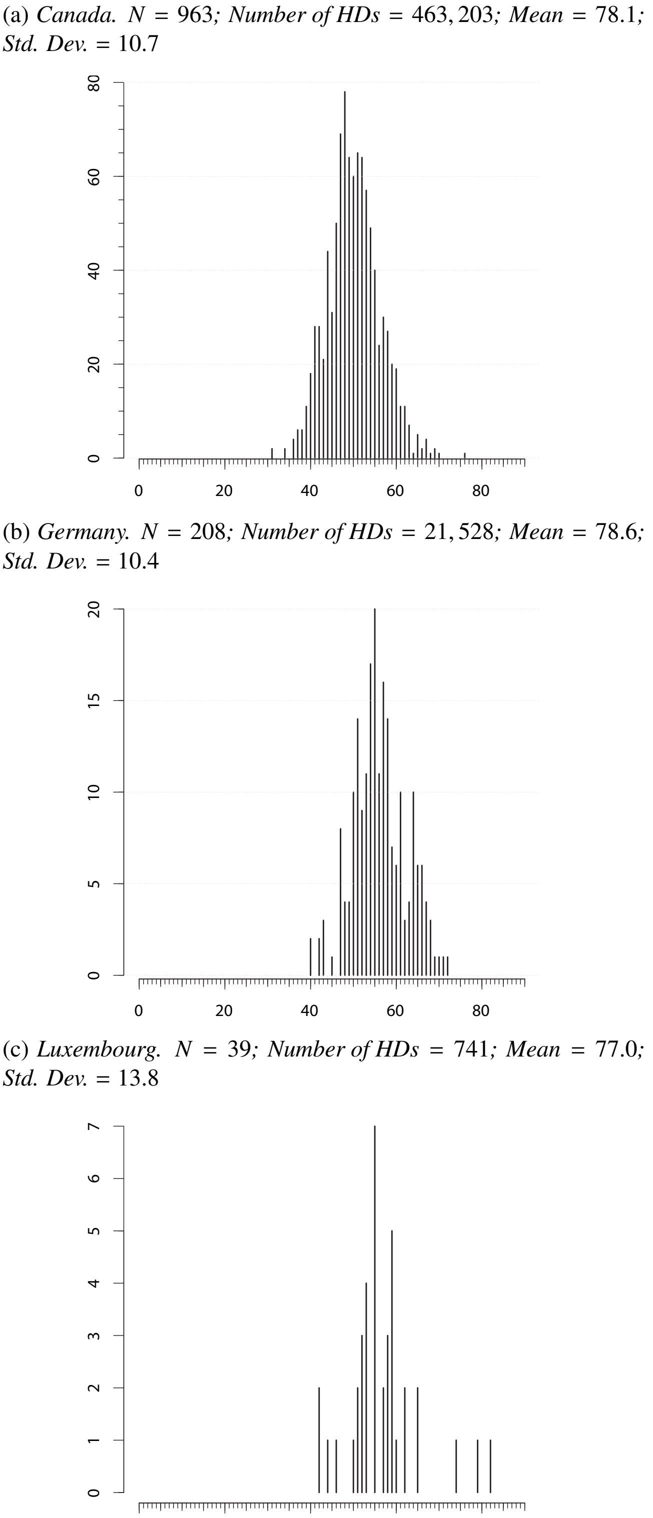
(a) *Italy. N* = 1038; *Number of HDs* = 538, 203; *Mean* = 70.8; *Std. Dev.* = 10.6

(b) *Slovenia. N* = 341; *Number of HDs* = 57, 970; *Mean* = 68.6; *Std. Dev.* = 10.7

(c) *United Arab Emirates. N* = 333; *Number of HDs* = 55, 278; *Mean* = 85.9; *Std. Dev.* = 17.5

Figure 5

Hamming distances, Q-Q-plots, selected suspicious countries

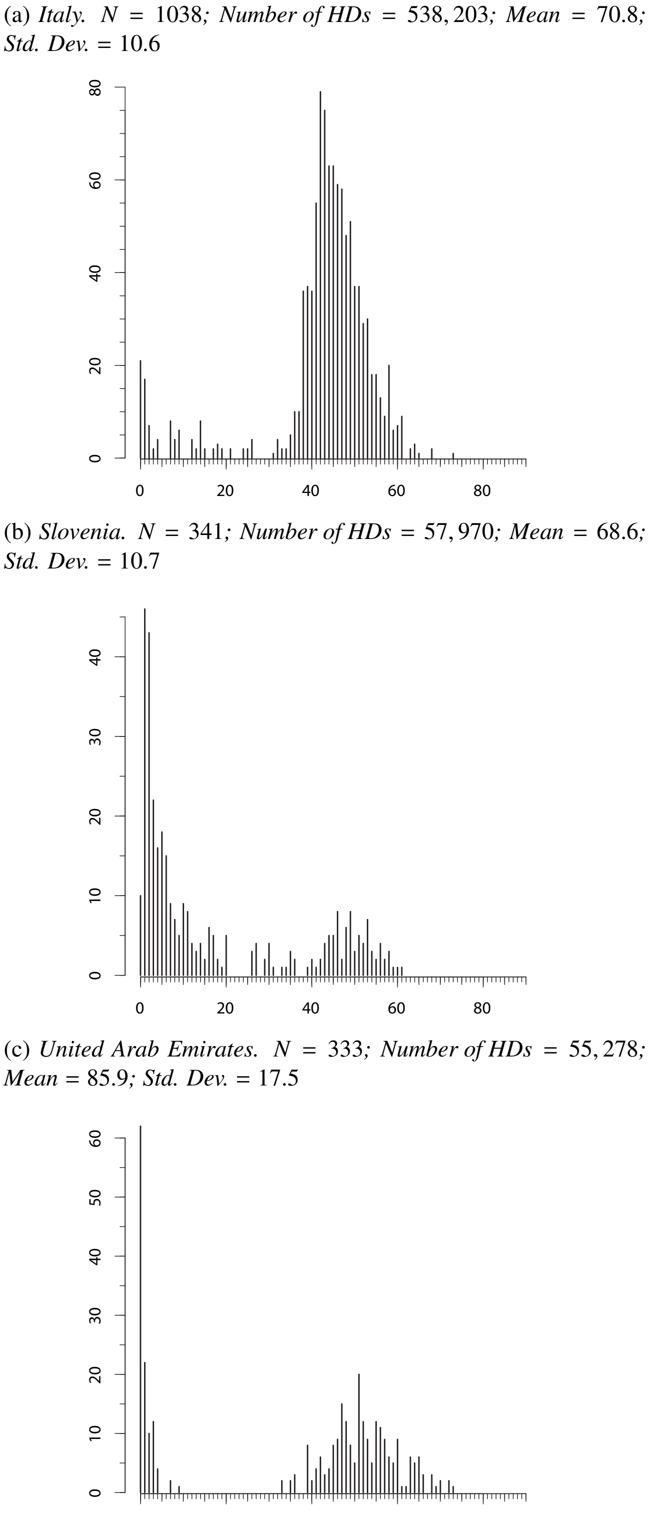

(a) *Canada. N* = 963*; Number of HDs* = 463, 203*; Mean* = 78.1*; Std. Dev.* = 10.7

(b) *Germany. N* = 208*; Number of HDs* = 21, 528*; Mean* = 78.6*; Std. Dev.* = 10.4

(c) *Luxembourg. N* = 39*; Number of HDs* = 741*; Mean* = 77.0*; Std. Dev.* = 13.8

Figure 6

Hamming distances, bar charts, minimum distances, selected unsuspicious countries


(a) *Italy. N* = 1038*; Number of HDs* = 538, 203*; Mean* = 70.8*; Std. Dev.* = 10.6

(b) *Slovenia. N* = 341*; Number of HDs* = 57, 970*; Mean* = 68.6*; Std. Dev.* = 10.7

(c) *United Arab Emirates. N* = 333*; Number of HDs* = 55, 278*; Mean* = 85.9*; Std. Dev.* = 17.5

Figure 7

Hamming distances, minimum distances, bar charts, selected suspicious countries

The aforementioned deviations from the normal distribution for Italy, Slovenia, and the UAE on the left-hand side become apparent in the Q-Q plots (Figure 5). At the lower left edge of the small quantiles, where IRPs and NIRPs are observable (cf. Table 3), the empirical frequencies deviate strongly from the expected distribution. Excepting the large deviations in the left part of the figures, reflecting IRPs and NIRPs, all three countries follow, for most of their HDs, the expected frequency distribution.

The Q-Q plots for all countries in the PISA 2009 datasets are shown in the Online Appendix (Figure B). Except for a few countries, the observed quantiles are comparable to the normal distribution. Besides the three countries with suspicious data mentioned above, other countries which exhibit some outlying HDs include Austria, Belgium, Colombia, Hungary, Montenegro, Switzerland, and Uruguay. Although the number of outlying HDs is small, it is very likely that some kind of mistake has occurred; the respective interviews should be deleted.

The following plots show country-wise bar charts for the *smallest* HD ($\mathrm{HD}_{\mathrm{min}}$) that are found for each case in the data. The numbers of the visualized distances in these plots are equal to the sample size. The minimum distances also follow the normal distribution when the sample size is sufficient, as is true for Canada. Although Kuriakose and Robbins (2016) also compute the minimum distances, using the Hamming distance there is no need for a threshold value, as outlying HDs can be detected via their distance to the mean, to be measured in standard deviations. In other words, using the HD it is possible to compute the probability of occurrence of each NIRP. The bar charts for Canada, Germany, and Luxembourg (Figure 6) show that no case in these countries has an $\mathrm{HD}_{\mathrm{min}} \leq 30$.

In contrast to the unsuspicious countries shown in Figure 6, Italy, Slovenia, and the UAE again give very different pictures (Figure 7). In Italy, the majority of cases has an $\mathrm{HD}_{\mathrm{min}} > 30$, but there is a significant number of cases with an $\mathrm{HD} \leq 30$. As Table 4 shows, 21 cases have at least one identical counterpart, and can thus be considered IRPs. The difference to the 12 IRPs shown in Table 3 results from the fact that IRPs involving exactly two cases are counted twice in Table 4, i.e., if case A is identical to B, both cases have $\mathrm{HD}_{\mathrm{min}} = 0$; in Table 3 they were counted as one pair of cases, the distance between them being $\mathrm{HD} = 0$. If there are three identical cases, they count as three HDs in Table 3 (cases A, B, C, comprising pairs AB, AC, and BC, all of them with $\mathrm{HD} = 0$) and count three times in Table 4 (cases A, B, and C, all of them with $\mathrm{HD}_{\mathrm{min}} = 0$). In the given example there are nine duplicates and one triplet (with respect to $\mathrm{HD}_{\mathrm{min}}$, nine duplicates are equal to 18 cases, one triplet consists of three cases, ergo in sum there are 21 cases with distances of $\mathrm{HD}_{\mathrm{min}} = 0$; with respect to HD, nine duplicates add up to nine pair-comparisons with $\mathrm{HD} = 0$, one triplet to three pair-comparisons with $\mathrm{HD} = 0$, resulting in 12 pair-comparisons with $\mathrm{HD} = 0$). In addition, 17 cases have $\mathrm{HD}_{\mathrm{min}} = 1$, seven cases have $\mathrm{HD}_{\mathrm{min}} = 2$, two cases have a $\mathrm{HD}_{\mathrm{min}} = 3$, and four cases have $\mathrm{HD}_{\mathrm{min}} = 4$. In total, there are 69 cases with $\mathrm{HD}_{\mathrm{min}} \leq 10$. For the UAE, Table 3 shows 31 pair-comparisons of $\mathrm{HD} = 0$, while Table 4 shows 62 cases with $\mathrm{HD}_{\mathrm{min}} = 0$. It can be concluded that the responses of 31 principals have been duplicated without further modification, resulting in the responses of 62 ostensible principals. The solutions for all countries are shown in the Online Appendix (Figure C).

6 Discussion

Although there is considerable discussion of interviewer effects and the prevention of data fabrication by interviewers (Blasius & Thiessen, 2021; Crespi, 1945; Groves, 2004; Olson et al., 2020), there is very little discussion of how to detect IRPs and NIRPs in published datasets (Blasius, 2018; Blasius & Thiessen, 2012, 2015; Hernandez et al., 2022; Koczela et al., 2015; Kuriakose & Robbins, 2016; Sarracino & Mikucka, 2017; Slomczynski et al., 2017). Kuriakose and Robbins (2016) describe duplicated data, and identify survey companies as a possible source for fabricated data. To our knowledge, Blasius and Thiessen (2015) and Blasius (2018) were the first to propose employees of survey research organizations as the origin for data fabrication. Because the PISA principal data were collected without the presence of an interviewer, only the employees of survey research organizations with access to the electronic data file could have duplicated these data.

The string-based technique proposed here provides a quick and easy way to detect copy-and-paste strategies in survey data. As shown, the frequency distributions of all distances from the entire set of pair-by-pair comparisons and HDs, as well as the minimum HDs, can be used to identify IRPs and NIRPs in the dataset.

In comparison to existing methods, our approach offers two advantages: First, in contrast to Blasius and Thiessen (2015), Slomczynski et al. (2017), and Blasius (2018), the HD enables the detection of NIRPs—it is possible to show both how many variable categories in each pair of cases are different and the probability of their occurrence. Applying MCA or PCA allows researchers to find IRPs (duplicates, triplets, . . .); in this case there is no difference to HD. If the number of variables and cases is strongly restricted, and if one is willing to compare factor scores for a number of dimensions, it is possible to find NIRPs with a limited number of differences. In contrast, when applying the HD, all IRPs and all NIRPs on all levels can be computed. The analysis of string distances is restricted only by computational power, but even a cross-comparison of 18,000 cases and 200 variables (see Figure 1) can be performed with a standard laptop.

The most important advantage of the proposed HD ap-

Table 4

Minimum Hamming distances, shown are the frequencies for $HD_{min} = 0$ to $HD_{min} = 20$, selected countries

Distance	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Sum
Australia	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Austria	4	0	4	0	0	2	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14
Belgium	2	2	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	2	0	0	0	8
Colombia	4	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
Czech Republic	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Hungary	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Italy	21	17	7	2	4	0	0	8	4	6	0	0	4	2	8	2	0	2	3	2	0	92
Latvia	2	0	1	0	0	0	0	2	0	0	2	0	0	0	0	0	0	0	0	0	0	7
Mexico	6	0	0	0	0	0	2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	10
Netherlands	4	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	6
Qatar	4	2	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
Slovak Republic	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Slovenia	10	46	43	22	16	18	15	9	7	5	9	8	4	3	4	2	6	5	2	1	5	240
Spain	8	0	2	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14
Switzerland	4	0	0	0	2	0	3	0	0	0	0	0	0	0	2	0	0	0	0	0	0	11
Trinidad and Tobago	2	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	4
United Arab Emirates	62	22	10	12	4	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	113
Uruguay	2	0	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	2	0	0	0	8
Other Countries	2	0	2	2	0	0	0	0	2	0	0	2	0	0	0	0	2	0	2	2	0	16
Total	143	89	71	40	28	24	26	21	15	16	15	10	10	5	14	4	8	11	7	5	5	567

proach is that the entire set of all similarities between all cases is computed. From the distribution of all HDs within a dataset, the mean and standard deviation can be calculated, so that the probabilities of the occurrence of IRPs and NIRPs can be determined. These calculations include the number of variables and the number of their categories, the frequency distribution of the category values within the variables, the number of cases, and the homogeneity of the sample. It is not necessary to specify a fixed threshold for outlying similarities and to specify a certain percentage of cases in order to flag a dataset as (partly) fabricated. Based on the probabilities of the outlying HDs being part of the dataset, one can decide to either delete the respective cases and use the remaining part of the data, or to classify the entire dataset as fabricated. To pre-empt possible mistakes – for example, entering the same questionnaire data twice—we suggest that survey research organizations run the proposed procedure before publishing their data.

As shown by Sarracino and Mikucka (2017), estimates can already be biased even when the number of duplicated cases is relatively low. When deciding how to handle fabricated data, there are at least two scenarios that have to be distinguished from one another. In the first scenario, only a few interviews are duplicated, as shown for Austria, Belgium, Switzerland, and some other countries. In this case, one can assume this happened by accident. Since entire interviews are duplicated, for secondary analysis there is no way to establish which interview is the original and which is the copy. Therefore, our recommendation is to delete the interviews in question. In the second scenario, a large number of interviews have been duplicated, as is the case for Slovenia and the UAE. One can identify the respective interviews, and might then simply delete them, but the level of widespread manipulation in these countries gives rise to the following question: Can the remaining interviews be trusted when it is evident that at least a (relatively) large part of the data set has been fabricated?

In conclusion, an approach in which IRPs and NIRPs are visualized using the Hamming distance enables new ways of assessing data fabrication. Using this string distance, we were able to demonstrate that individual aberrant behaviour by employees of survey research organizations occurs in a very prominent survey: PISA data, which is gathered under the aegis of the OECD, has been used in thousands of studies all over the world, including publications in leading international journals. Although the findings presented here expose a shocking number of fabricated interviews—at least in Slovenia, Italy, and the UAE—we still believe PISA data to be among the best survey data that is currently publicly available.

References

Blasius, J. (2018). Fabrication of interview data. *Quality Assurance in Education*, *26*(2), 213–226.

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

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

Blasius, J., & Thiessen, V. (2021). Perceived corruption, trust, and interviewer behavior in 26 european countries. *Sociological Methods & Research*, *50*(2), 740–777.

Cohen, M. J., & Warner, Z. (2021). How to get better survey data more efficiently. *Political Analysis*, *29*(2), 121–138.

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

Gifi, A. (1990). *Nonlinear multivariate analysis*. Wiley.

Groves, B. (2004). Interviewer falsification in survey research: Current best methods for prevention, detection, and repair of its effects. *Survey Research*, *35*(1), 1–15.

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell System Technical Journal*, *29*(2), 147–160.

Hernandez, I., Ristow, T., & Hauenstein, M. (2022). Curbing curbstoning: Distributional methods to detect survey data fabrication by third-parties. *Psychological Methods*, *27*(1), 99–120.

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.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, *51*(2), 201–219.

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, *1996*(70), 29–44.

Kuriakose, N. L. (2015). *Percentmatch: Stata module to calculate the highest percentage match (near duplicates) between observations* [Statistical Software Components, Boston College Department of Economics].

Kuriakose, N. L., & Robbins, M. (2016). Don't get duped: Fraud through duplication in public opinion surveys. *Statistical Journal of the IAOS*, *32*(3), 283–291.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys (CSUR)*, *33*(1), 31–88.

OECD. (2009). Data base PISA 2009. https://www.oecd.org/pisa/data/pisa2009database-downloadabledata.htm

OECD. (2012). PISA 2009 technical report.

Olson, K., Smyth, J. D., Dykema, J., Holbrook, A. L., Kreuter, F., & West, B. T. (Eds.). (2020). *Interviewer effects from a total survey error perspective*. CRC Press.

Powałko, P. (2015). Duplicates in social-science surveys [Presentation on the 6th Conference of the European Survey Research Association, Reykjavik, July 13–17.]. https://www.europeansurveyresearch.org/conf2015/uploads/678/1645/160/ESRA_2015_Przemek_Powalko_Duplicates_in_Social_Science_Surveys.pdf

Sarracino, F., & Mikucka, M. (2017). Bias and efficiency loss in regression estimates due to duplicated observations: A monte carlo simulation. *Survey Research Methods*, *11*(1), 17–44.

Schafer, C., Schrapler, J.-P., Muller, K.-R., & Wagner, G. G. (2005). Automatic identification of faked and fraudulent interviews in the German SOEP. *Schmollers Jahrbuch*, *125*, 183–193.

Simmons, K., Mercer, A., Schwarzer, S., & Kennedy, C. (2016). Evaluating a new proposal for detecting data falsification in surveys. *Statistical Journal of the IAOS*, *32*(3), 327–338.

Slomczynski, K. M., Powalko, P., & Krauze, T. (2017). Non-unique records in international survey projects: The need for extending data quality control. *Survey Research Methods*, *11*(1), 1–16.

Van der Loo, M. P., et al. (2014). The stringdist package for approximate string matching. *The R Journal*, *6*(1), 111.

Yamamoto, K., & Lennon, M. L. (2018). Understanding and detecting data fabrication in large-scale assessments. *Quality Assurance in Education*, *26*(2), 196–212.