doi.org/10.26398/IJAS.0029-007

THE DIRTY DATA INDEX – ASSESSING THE QUALITY OF SURVEY DATA IN INTERNATIONAL COMPARISON

Jörg Blasius[1]

University of Bonn, Germany

Oleg Nenadić

University of Göttingen, Germany

Victor Thiessen

Dalhousie University, Halifax, Canada

Abstract*. In 2012, Blasius and Thiessen developed the dirty data index (DDI) as an index for measuring the quality of survey data. The DDI is based on the quantifications from categorical principal component analysis which works on item batteries of ordered categorical data such as five-point Likert-scaled items. This is an ongoing work where we further develop the index for application in international comparison. As an example we use data from the International Social Survey Programme 2012, Family and Changing Gender Roles, including 36 countries with a total of more than 56,000 cases.*

Keywords*: Data quality, Categorical principal component analysis, Ordered categorical data, International comparison.*

1. INTRODUCTION

When treating data, the basic assumption is that the data were carefully collected and that one can trust the numbers. Be it in the natural sciences, in economics, in medicine or in the social sciences - any computation and any result is likely to be wrong if the data quality is low. In experimental settings, as they are common in the natural sciences, psychology, and medicine, data have to be collected at a level of technical and methodological standard as high as possible to provide "proper solutions". In the social sciences, researchers usually work with survey data, a kind of data that is important for any society, for example, as the basis for political and economic decisions. As it is true for data from the natural sciences, psychology and

[1] Corresponding author: Jörg Blasius, email: jblasius@uni-bonn.de

medicine, survey data also have to be on as high as possible standards. However, these data are often far away from standards that are acceptable for any analysis; in many cases they are not collected in an appropriate manner. For example, the questionnaires may not be well formulated (Saris and Gallhofer, 2014) or the respondents did not take care when responding to the questions (Krosnick, 1991, 1999). Sometimes, survey data even comprise fabricated data either from the interviewers (Crespi, 1945, Blasius and Thiessen, 2012, 2013, Thiessen and Blasius, 2016) or from employees of the field agencies who collected the data (Blasius and Thiessen, 2012, 2015, Thiessen and Blasius, 2016). If the data are inappropriately collected or even fabricated, the quality of the data is compromised. Hence, these data should not be used for any analyses from which substantive solutions are drawn.

All survey data consist of a mixture of substantive and non-substantive variations (Blasius and Thiessen, 2012). Thereby, the substantive variation is the part researchers are interested in for explaining phenomena in the society. Non-substantive variations come from numerous sources such as socially desirable responding, response styles, failure to understand questions, and from fabricated interviews.

There is a large discussion on the reasons for poor data quality, especially connected with the term "satisficing", which was first introduced by Simon (1957) to situations where humans do not strive for an optimization of outcomes. Krosnick (1991, 1999) recognized that the survey setting typically induces satisficing that can take place at different stages through data collection and simply means less careful fulfillment of answering the survey questions. Blasius and Thiessen (2012) introduced the term of 'simplification', since it includes all strategies and all sources aimed at reducing the efforts in answering the questions; simplification can also be done by interviewers when fabricating (parts of) their interviews (Crespi, 1945, Blasius and Thiessen, 2012, 2013) and employees of field agencies, for example, via copy-and-paste procedures (Blasius and Thiessen, 2012, 2015).

As discussed by Blasius and Thiessen (2012), even the most well-known and most frequently used survey data, for example, the International Social Survey Programme (ISSP), the World Value Survey, and PISA contain at least parts of poor quality. It follows that any subsequent analyses will be of equally poor quality unless steps are taken to detect, isolate, and take account of the artefactual variations. In this paper, we introduce a method for the detection of low quality data that is based on response styles, misunderstanding of questions because of poor item construction, heterogeneous understanding of questions arising from cultural differences, or faked or partly faked interviews. In the following, we restrict our

attention to the underlying structure of responses to a set of statements on a particular topic and to ordered categorical data as given by Likert-scaled items. Whatever are the reasons for low data quality, in this paper we further develop the dirty data index (DDI) for measuring the quality of data. As an example, we use data from the ISSP 2012, Family and Changing Gender Roles.

2. DATA QUALITY AND ORDERED CATEGORICAL DATA

Ordered categorical data such as five-point-items are often treated as metric, suggesting that there are equal distances between the categories. Considering an item with five response categories running from "strongly agree" via "agree", "neither agree nor disagree", "disagree" to "strongly disagree", numbered with equal distances from "1" to "5"; in contrast to the measurement level, the latent distances between the categories differ, the distance from "strongly agree" to "neither agree nor disagree" is not twice as long as the distance to "agree" or the distance between "agree" and "neither agree nor disagree". The latent distances will also change in case the categories are labelled somewhat differently, for example, the second category is labelled with "agree somewhat" and/or the first category with "agree". Moreover, the distances on the positive part of the scale might be different from those on the negative part of the scale, even when the items are constructed symmetrically. Furthermore, it depends on the exact wording of the questions. Small changes in a statement such as adding/deleting/changing a single word may change the respondents' answer from one category to the adjacent (or to a farther one). Take the question "A job is all right, but what most women really want is a home and children" as an example and change it to "A job is all right, but what all women really want is a home and children" – one may "strongly agree" with the term "most women" but not with "all women". Both questions are very similar but the margins of their categories are different and therefore the latent distances between the categories are also different.

A further problem in the data is connected with the middle category "neither agree nor disagree". Depending on the difficulty of this question, this category is also used as an alternative to "I don't know" (Blasius and Thiessen, 2001). If somebody is fully aware of the question's topic, after thoughtful consideration of all aspects s/he might come to the conclusion "neither nor" is the true answer. Others might not have spent a single thought onto the topic of this question but prefer to hide this fact, in this case the given response is "neither nor" instead of "I don't know". The answer of the question also depends on the cultural background, i.e., in the given context of how important family is in the country– and how strong the

government supports families, e.g., via financial support, nursery schools, and job offers for women with (small) children. Finally, the answer to the question depends on the social status of the respondents, for somebody coming from the working class it is rarely a question "what most or all women really want", the women have to work for financial reasons.

To summarize, it depends on the context, on the wording, and on the cultural background how to answer a single question. It is also a matter of education whether respondents understand the context of the question; also in well-known surveys we found questions that were hardly understandable for respondents with low education or low interest in the topic (Blasius and Thiessen, 2001, 2012).

3. SCALING ORDERED CATEGORICAL DATA

In survey research, many items contain an implicit order for adjacent categories of survey items, such as those with a Likert response format, which one may wish to impose on the data. In such situations one expects that a "strongly agree" response implies greater agreement with a given statement than an "agree" response, which in turn should reflect a stronger agreement than a "neither agree nor disagree" and so on. The lowest agreement or the highest disagreement should be indicated by "strongly disagree" responses. In contrast to scaling techniques such as multiple correspondence analysis, categorical principal component analysis (CatPCA) permits the assumed order of the successive categories to be tested, and in contrast to scaling techniques such as principal component analysis (PCA) and factor analysis, it does not require equal distances between successive categories.

In CatPCA, the categories of the original items are replaced by optimal scores, the quantification values (cf. Gifi, 1990). The optimal scoring process allows order constraints to be imposed so that ordered categorical variables increase, or at least do not decrease, with increasing category codes. Responses that are inconsistent with the implied ordering, manifest themselves in tied optimal scores for two or more successive categories.

In contrast to PCA, the number of dimensions (m^*) to be considered in the solution must be specified in advance, and the solutions for m^* and $m^* + 1$ dimensions are not nested. Once the optimal scores have been calculated, they replace the category codes and the remainder of the analysis can be regarded as (classical) PCA. In short, CatPCA is an appropriate technique to display relationships between cases associated with a set of ordered categorical variables.

Like PCA, CatPCA produces eigenvalues and explained variances, factor loadings, and factor scores with mean zero and unit standard deviation. To

summarize, CatPCA can be regarded as PCA applied to ordered categorical data. For both methods, the aim is to approximate the elements of the matrix **Z** by the product of the factor scores times the factor loadings within the m^*-dimensional space, $\mathbf{Z} \approx \mathbf{FA}$. The degree of approximation can be measured by a least-squares loss function (cf. Gifi, 1990, de Leeuw, 2014):

PCA is commonly understood as a linear technique, since observed variables are approximated by linear combinations of principal components. However, it is possible to understand PCA as a bilinear technique, since the elements of the data matrix **Z** are approximated by inner products that are bilinear functions of factor scores **F** and factor loadings **A**.

CatPCA can be seen as nonlinear transformations of the variables that still preserve the (bi)linearity of the technique. It follows that loss is not merely minimized over the factor scores and factor loadings, but also over the admissible transformations of the columns of **Z** (for further details of the method, see Gifi, 1990, de Leeuw, 2006).

4. DATA

Cross-national research is one of the most prominent topics in the social sciences. Since the 1970s, different survey programs have been designed explicitly for cross-national comparisons; the most renowned ones include the World Values Surveys, The European Social Surveys, the Program for International Student Assessment (PISA), and the International Social Survey Program (ISSP). The ISSP data alone which are reputed to maintain high standards (Scheuch, 2000), have been used in several thousand research papers, which are listed on 328 pages in nine-point font (see http://issp.org/fileadmin/user_upload/Bibliography/2017_ISSP_Bibliography.pdf; access of December 18, 2017). In general, the ISSP can be seen as one of the best international surveys. Currently, 45 countries from all continents participate (while not every country participates every year) with sample sizes exceeding 1,000 interviews in each country in each year.

We use the 2012 data focusing on "Family and Changing Gender Roles" with 36 countries participating (excluding Spain, because they use a four-point instead of a five-point scale). For our analyses, we use seven variables with five response categories (from "strongly agree" to "strongly disagree") measuring the disposition to working mothers (for the wording of the variables, see Table 1). To show parts of the data, we provide the distributions for the seven variables and for three countries: Australia, The Philippines, and the USA, excluding all cases with one or more missing responses (Table 1), which reduces the data sets from 1.612 to 1.432

(Australia), from 1.200 to 1.177 (The Philippines), and from 1.302 to 1.137 (USA). The proportion of missing values is often used as an indicator of the data quality (Biemer and Lyberg, 2003, West and Blom, 2016); according to this criterion The Philippines should have the best data. Without discussing details of the table, it is evident that the answer structures differ strongly by country. Please note that The Philippines have very high numbers on "strongly agree" in all questions, and that some questions may have no straightforward answers, for example, "Both the man and the woman should contribute to the household income" – if one replaces the word "should" by "have to" the meaning is very different. For some countries, this is probably the more appropriate formulation since for many respondents it is not a question of "should", but the women have to work.

Table 1: Distribution of *disposition to working mothers* items for three selected countries: Australia (AU): *N* = 1,432; The Philippines (PH): *N* = 1,177; USA: *N* = 1,137 (listwise deletion)

Item	C	SA	AS	NN	DS	SD	χ^2
A working mother can establish just as warm and secure a relationship with her children as a mother who does not work.	AU	23.0	45.3	11.0	17.2	3.6	
	PH	37.8	34.2	13.0	12.5	2.5	244.3
	USA	27.4	44.5	22.6	5.5	0.0	
A pre-school child is likely to suffer if his or her mother works.	AU	5.0	25.8	18.6	36.7	13.9	
	PH	29.7	35.4	13.3	18.4	3.2	1,097.0
	USA	6.7	26.7	50.7	15.8	0.0	
All in all, family life suffers when the woman has a full-time job.	AU	6.7	28.9	17.5	30.5	16.4	
	PH	21.4	27.5	18.4	26.7	6.0	320.2
	USA	5.5	22.7	12.0	39.4	20.4	
A job is all right, but what most women really want is a home and children.	AU	5.7	21.4	25.6	31.7	15.6	
	PH	29.2	37.6	15.5	13.1	4.5	605.0
	USA	6.2	29.0	24.5	30.5	9.8	
Being a housewife is just as fulfilling as working for pay.	AU	12.6	37.0	26.5	18.4	5.4	
	PH	27.4	45.0	16.1	8.9	2.6	205.1
	USA	14.3	43.4	17.9	19.6	4.7	
Both the man and the woman should contribute to the household income.	AU	14.2	38.4	31.8	13.7	2.0	
	PH	62.3	29.5	4.9	2.5	0.8	924.5
	USA	19.3	45.9	24.3	9.1	1.4	
A man's job is to earn money; a woman's job is to look after the home and the family.	AU	2.9	13.1	19.8	38.0	26.2	
	PH	48.7	31.5	7.6	10.1	2.0	1,576.9
	USA	4.7	17.6	17.4	42.8	17.4	

SA=Strongly agree, AS=Agree somewhat, NN=Neither agree nor disagree, DS=Disagree somewhat, SD=Strongly disagree

For assessing the quality of the responses, we first scale the items using CatPCA and study the quantifications. As described by Blasius and Gower (2005), the more similar the distances between categories on the latent scale, the better the

related categories differ in the questions – otherwise, the more dissimilar the distances between categories on the latent scale, the worse the successive categories can be distinguished. In the extreme case we observe ties on the latent scale, which means that there is no differentiation between the respective categories. As Table 2 shows, Australia has no ties at all, in the USA in the fifth question, the last three categories are tied and in the sixth questions even the last four categories are tied. The quantifications for The Philippines show the same pattern for all questions with the last four categories being tied. In other words, The Philippines either strongly agreed with the statement or they chose any other of the four response options. This is a clear indicator for low data quality which we found also for other data for the Philippines (for example, Blasius and Thiessen, 2006, 2009).

Table 2: CatPCA quantifications by selected countries, two-dimensional-solution

Item	C	SA	AS	NN	DS	SD	Ties
A working mother can establish just as warm and secure a relationship with her children as a mother who does not work.	AU	-1.367	-0.201	0.825	1.460	1.768	None
	PH	-1.283	0.780	0.780	0.780	0.780	0.454
	USA	-1.364	0.014	1.242	1.606	—	None
A pre-school child is likely to suffer if his or her mother works.	AU	-1.474	-1.146	-0.344	0.509	1.771	None
	PH	-1.537	0.651	0.651	0.651	0.651	0.596
	USA	-1.493	-1.157	0.278	1.694	—	None
All in all, family life suffers when the woman has a full-time job.	AU	-1.427	-1.014	-0.222	0.473	1.727	None
	PH	-1.916	0.522	0.522	0.522	0.522	0.898
	USA	-1.490	-1.243	-0.470	0.280	1.519	None
A job is all right, but what most women really want is a home and children.	AU	-1.555	-1.042	-0.468	0.485	1.789	None
	PH	-1.556	0.643	0.643	0.643	0.643	0.552
	USA	-3.105	-0.690	0.324	0.753	0.843	None
Being a housewife is just as fulfilling as working for pay.	AU	-1.397	-0.646	0.312	1.035	2.643	None
	PH	-1.630	0.614	0.614	0.614	0.614	0.418
	USA	-2.422	0.259	0.555	0.555	0.555	0.468
Both the man and the woman should contribute to the household income.	AU	-1.819	-0.391	0.505	1.570	1.664	None
	PH	-0.778	1.285	1.285	1.285	1.285	0.124
	USA	-2.047	0.488	0.488	0.488	0.488	0.291
A man's job is to earn money; a woman's job is to look after the home and the family.	AU	-1.791	-1.499	-0.743	0.089	1.382	None
	PH	-1.027	0.974	0.974	0.974	0.974	0.340
	USA	-2.844	-1.221	-0.316	0.464	1.184	None

In a first attempt to measure the quality of survey data, Blasius and Thiessen (2009) counted the number of respondents involved in tied data and showed that they can be used as a rough indicator for data quality. If two categories were tied, they took the smaller number of respondents included in the tied data, if three categories were tied, they multiplied the lower number of respondents which were

tied by two and added them to the number of respondents already considered in the bridging category, in case four categories were tied, they multiplied the minimum number of respondents by three and went on as described for the case with three tied categories. Table 2 also shows the results of this procedure, for a better comparison between the countries we standardized resulting values by the number of cases. To give a computation example, we used the second item from The Philippines: A pre-school child is likely to suffer if his or her mother works (cf. Table 1; given are the percentages, for the following calculation we need the frequencies). The numbers of respondents for the four tied categories are 416 (agree; $.354 \times 1.177 = 416$), 156 (neither nor), 216 (disagree), and 38 (strongly disagree). From the two categories with the largest distance on the original data (agree and strongly disagree), the category "strongly disagree" has the smaller number of cases, it will be multiplied by "3", the subsequent category by "2" and the last one by "1". It follows: *N of ties*, second item The Philippines = $(3 \times 38 + 2 \times 216 + 156) = 702$, divided by the number of cases, e.g. $702 / 1177 = 0.596$. It is evident that the more ties are in the data and the more cases are involved, the higher the value becomes. The highest value in the table is 0.898; it belongs to The Philippines and to the question "All in all, family life suffers when the woman has a full-time job". The answers to this question seem to be arbitrary for a large number of respondents. But indeed, the question is not well formulated, for many respondents from families with full-time working women it might be just a matter of fact that "family life suffers when the woman has a full-time job", they may not see the intention of the researchers when formulating this question. A strong limitation of this measure is that it just subdivides between ties and non-ties – the latent distances between the categories are not considered. This will be done in the next step by introducing the DDI.

5. AN IDEA FOR MEASURING THE QUALITY OF ORDERED CATEGORICAL DATA

CatPCA quantification values are standardized to mean of zero and standard deviation of one (Gifi, 1990). The smaller the differences between PCA and CatPCA solutions, the closer is the distribution of the quantification values to the standard normal distribution and the higher is the data quality (Blasius and Gower, 2005). Based on the features of CatPCA, Blasius and Thiessen (2012) created an index to measure the quality of survey data. The index uses the quantification values of CatPCA to compute the respective probability area from the standard normal distribution and to compare this probability area subsequently with the probability area obtained from the frequencies. If, for example, there is a total of 1,000 cases

with 100 belonging to the first category, the left hand area of the standard normal distribution would contain 10% of the cases with a corresponding *z*-value of -1.28. The midpoint would contain 2.5% of the area with a respective *z*-value of -1.96. Moving from a *z*-value to the proportion of cases, a *z*-value of -1.0 would require 15.87% of the cases to the left and 84.13% cases to the right, while a *z*-value of 0.0 would divide the number of cases exactly in the middle.

Comparing the probability areas received from the marginal of each item with the probability areas one can compute from the CatPCA quantifications, a value can be computed for each item category showing how close the quantification value is to the expected value. Performing such calculations for all categories of an item, counting them, and dividing them by the number of categories, one obtains a value that indicates the quality of the item in the given set of questions. These calculations are done for all items in the item battery, and the mean of these values is called "dirty data index" (DDI). The lower bound of this value – and also of the value of a single item – is zero, which is not obtainable in survey research since ordered categorical data do not have equal distances. From simulation studies, Blasius and Thiessen (2012) calculated $\frac{k}{k-1}$, with k = number of categories, as upper bound of the value. However, values greater than one are very rare in survey data.

The procedure is visualized in Figure 1 where we exemplified the calculations for the first question of the Family and Changing Gender Roles item battery from the ISSP 2012 data for Australia: "A working mother can establish just as warm and secure a relationship with her children as a mother who does not work." The first category (strongly agree) contains 329 cases (or 23.0%) of the 1,432 cases without missing values on the entire item battery. The corresponding *z*-value is -0.74 and the 23.0 % area under the standard normal distribution is symbolized by the left-most line with small dashes. The *z*-value for the midpoint of this area (11.5% to the left) is *z* = -1.20, symbolized by long dashes. The CatPCA quantification value for the first category of the item is -1.37, shown as a solid line, and its respective quantification area is 0.0859 (cf. Table 3). The difference between the area from the category midpoint and the area computed from the CatPCA quantification value is 0.1149 – 0.0859 = 0.0290; this area is shaded in Figure 1. The second category contains 45.3% of the cases, its midpoint is 23.0 + 45.3/2 = 45.7 with a *z*-value of -0.20; the respective CatPCA quantification value is -0.2008 (cf. Table 3). The area between the midpoint from the original data and the quantification area is 0.0356 (also shaded); the values for the three other areas are given in Table 1 and shadowed in Figure 1. After computing the areas for all five categories, the differences between the areas from the normal distribution and from the

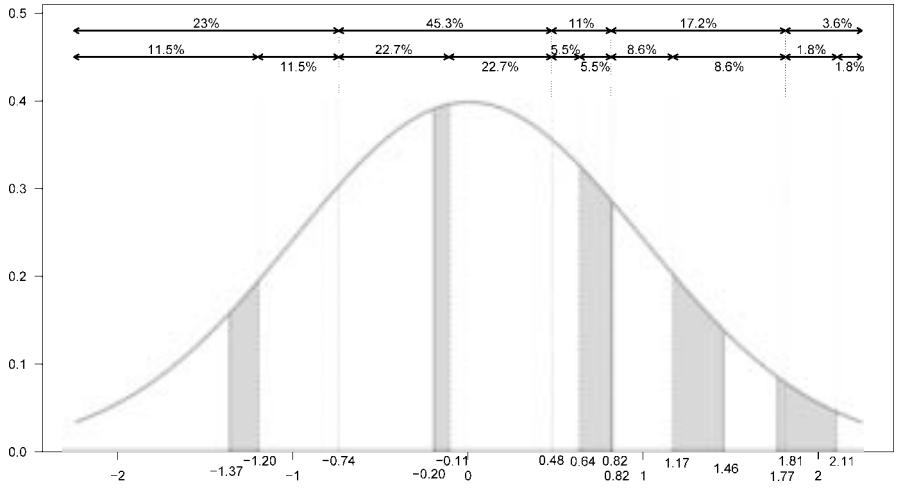

Figure 1: First item (Australia): "A working mother can establish just as warm and secure..."

quantifications are summed up, resulting in the DDI for the first item. Summing up the values for all items and dividing the resulting value by the number of variables gives the DDI.

In the following we explain the computational solution step by step (see also Blasius and Thiessen, 2012: 135-136), in which N = number of cases, K = number of items, with k = item index, J_K = number of categories in each item, with j = item category index, f_{jk} = frequency of category j of item k, m_{jk} = mass (relative frequency of category j of item k), $c_{(1,j)k}$ = cumulative relative frequencies of categories of item k, q_{jk} = quantification of category j of item k (provided by CatPCA). It yields: relative frequencies (masses) for each category: $m_j = f_j / N$, and cumulative masses: $c_{(1,j)} = m_j + c_{(1,j-1)}$ (if $j = 1$, $c_{(1,j-1)} = 0$).

First step: Compute the midpoints of the item categories. Start with the relative frequency (mass) of the first category and divide its value by 2; $g_1 = m_1 / 2$ (for $j = 1$) $\rightarrow$ first midpoint of the relative frequencies. Add the mass of the first category (m_1) to half the mass of the second category ($m_2 / 2$); $g_2 = g_1 + m_2 / 2 \rightarrow$ second midpoint value. Add the masses of the first two masses plus half the mass of the third category, ..., add the first ($J_K - 1$) masses ($c_{(1,J-1)}$) plus half the mass of the last category. Note: The number of midpoints is the same as the number of categories (= J_K): Do: $j = 1$ to J_K (for each item k); $g_j = g_{j-1} + m_j / 2$ (with $g_0 = 0$).

Second step: Compute the areas to the left of the quantification values (q_{jk}) on the basis of the standard normal distribution $\Phi(k)$.

Third step: Compute the absolute differences between midpoint (areas) and quantification (areas) and add them.

Table 3: First item (Australia): A working mother can establish just as warm and secure ...

Cat.	A) Freq.	B) Quantif.	C) Mass	D) cum. Mass	E) Mpts Area	F) Qtf.	G) Diff.
SA	329	-1.3665	0.2297	0.2297	0.1149	0.0859	0.0290
AS	648	-0.2008	0.4525	0.6823	0.4560	0.4204	0.0356
NN	158	0.8248	0.1103	0.7926	0.7374	0.7953	0.0578
DS	246	1.4604	0.1718	0.9644	0.8785	0.9279	0.0494
SD	51	1.7678	0.0356	1.0000	0.9822	0.9615	0.0207
Sum	1,432		1.0000				0.1926

With C(1) = A(1) / *N* = 329 / 1,432 = 0.2297; D = cumulative values of C;
E(1) = D(1) / 2 = 0.2297 / 2 = 0.1149;
E(2) = D(1) + C(2) / 2 = 0.2297 + 0.4525 / 2 = 0.4560;
…
E(5) = D(4) + C(5) / 2 = 0.9644 + 0.0356 / 2 = 0.9822
F(1) = Φ(B1) = Φ(-1.3665) = 0.0859
DDI(Q1_Australia): 0.0290 + 0.0356 + 0.578 + 0.0494 + 0.0207 = 0.1926.

When correcting this value by the upper bound as suggested by Blasius and Thiessen (2012) one receives: DDI(Q1_Australia, corrected): 0.1926 * 4 / 5 = 0.154.

In their simulation study, Blasius and Thiessen (2012: 136) showed that the DDI (corrected by the upper bound) fluctuates between 0.5 and 0.7 when using random data, depending on the given distribution ("normal", "U-shaped", …). They concluded that DDI values smaller than 0.3 indicate relatively good data and values smaller than 0.15 indicate data of exceptional quality. Vice versa, values greater than 0.5 indicate a low data quality.

6. SOLUTIONS

In the following, we apply the DDI to the 36 countries taking part in the ISSP 2012 using the seven items from the item battery of "Family and Changing Gender Roles". Table 4 shows the DDI solutions for all countries and all items without correcting them by the upper bound. It can be stated for each item in each country how well it has been understood by the respondents.

Table 4: DDI by countries (family and changing gender data), single questions, uncorrected values

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	DDI
Argentina	0.555	0.225	0.214	0.524	0.955	0.583	0.329	0.484
Australia	0.193	0.183	0.173	0.168	0.103	0.125	0.140	0.155
Austria	0.161	0.303	0.320	0.278	0.431	0.254	0.241	0.284
Bulgaria	0.488	0.615	0.713	0.425	0.514	0.180	0.290	0.460
Canada	0.274	0.085	0.153	0.158	0.435	0.241	0.120	0.209
Chile	0.694	0.219	0.265	0.413	0.855	0.929	0.603	0.568
China	0.969	0.790	0.859	0.329	0.643	0.793	0.505	0.698
Taiwan	0.334	0.326	0.169	0.603	0.993	0.990	0.126	0.505
Croatia	0.270	0.166	0.135	0.375	0.456	0.465	0.384	0.321
Czech Republic	0.181	0.210	0.151	0.099	0.234	0.260	0.225	0.194
Denmark	0.264	0.325	0.343	0.154	0.428	0.465	0.274	0.321
Finland	0.214	0.175	0.194	0.345	0.360	0.179	0.159	0.233
France	0.264	0.126	0.148	0.105	0.144	0.575	0.164	0.218
Germany	0.304	0.321	0.278	0.178	0.246	0.380	0.183	0.270
Iceland	0.271	0.381	0.396	0.454	1.046	0.891	0.374	0.545
India	0.831	0.823	0.891	0.500	0.681	0.875	0.553	0.736
Ireland	0.166	0.133	0.125	0.153	0.490	0.259	0.150	0.211
Israel	0.193	0.225	0.258	0.373	0.235	0.163	0.230	0.239
Japan	0.163	0.165	0.169	0.399	0.780	1.013	0.509	0.456
South Korea	0.758	0.363	0.200	0.624	0.651	0.860	0.660	0.588
Latvia	0.273	0.194	0.225	0.303	0.294	0.159	0.200	0.235
Lithuania	0.166	0.221	0.280	0.149	0.193	0.340	0.216	0.224
Mexico	0.851	0.770	0.760	0.760	0.913	0.703	0.389	0.735
Norway	0.171	0.220	0.261	0.300	0.333	0.223	0.296	0.258
Philippines	0.661	0.754	0.923	0.740	0.846	0.439	0.536	0.700
Poland	0.189	0.106	0.135	0.359	0.613	0.631	0.156	0.313
Russia	0.285	0.110	0.134	0.270	0.291	0.425	0.173	0.241
Slovakia	0.191	0.108	0.121	0.149	0.175	0.149	0.080	0.139
Slovenia	0.648	0.284	0.198	0.201	0.254	0.738	0.264	0.369
South Africa	0.819	0.421	0.359	0.431	0.468	0.683	0.388	0.510
Sweden	0.151	0.201	0.250	0.320	0.733	0.468	0.213	0.334
Switzerland	0.291	0.248	0.226	0.405	0.674	0.693	0.344	0.411
Turkey	0.935	0.081	0.276	0.208	0.168	0.721	0.799	0.455
Great Britainn	0.384	0.396	0.370	0.190	0.439	0.890	0.171	0.406
United States	0.138	0.156	0.174	0.396	0.760	0.985	0.210	0.403
Venezuela	0.943	0.918	0.856	0.599	0.559	0.794	0.556	0.746
Average	0.407	0.315	0.325	0.345	0.511	0.542	0.311	0.394

Table 4 also shows the average DDI value for each item that indicates how *well* the item has been understood across all countries under investigation. According to our calculations, questions five and six are the most problematic ones, followed by the first question. Especially the fifth question provides some extremely high values, for example, for Argentina, Taiwan, and Iceland; in the latter country the value even exceeds the threshold of 1.0. These high values may also depend on the culture of the country and on how to respond to an item such as "being a housewife is just as fulfilling as working for pay" – in some countries this statement might be widely understood as a joke rather than as a serious question. On country level, we find the highest values for China, India, Mexico, The Philippines, and Venezuela, e.g. countries that have been characterized in other papers as providing low data quality (for example, Blasius and Thiessen, 2006, 2012). According to the simulation from Blasius and Thiessen (2012), the values are even worse than random data.

In the next step, we run the same analyses while performing a three-dimensional solution in CatPCA (Table 5). Since three dimensions explain by definition more variation than two dimensions, the values of the DDI should decrease in all cases, at least on the level of the average values for the countries. For most countries, this assumption is true and the average values for the items are in all cases clearly smaller than in the two-dimensional solution. However, there are a few countries in which the DDI even slightly increases, i.e. South Korea (from 0.588 to 0.604), Latvia (from 0.235 to 0.276), and Slovakia (from 0.139 to 0.175), others remain on a very high level, for example, The Philippines (in both solutions 0.700), Mexico (from 0.735 to 0.725), and Turkey (from 0.455 to 0.444). For South Korea, The Philippines, and Mexico we can conclude that the answers are almost at random, which might also be caused by interviewers who do not take care of instructions or just fabricate the data (Blasius and Thiessen 2012, 2013). For Latvia and Slovakia, we assume that many respondents answered in a very simplified way, providing an almost one-dimensional solution (with a high value on Cronbach's alpha) resulting in an additional third factor that is just at random (cf. Blasius and Thiessen, 2012: chapter 8, Thiessen and Blasius, 2008).

Table 5: DDI by countries (family and changing gender data), single questions, three-dimensional CatPCA-solution

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	DDI
Argentina	0.368	0.165	0.099	0.364	0.461	0.266	0.301	0.289
Australia	0.196	0.171	0.161	0.098	0.064	0.189	0.111	0.141
Austria	0.173	0.320	0.350	0.176	0.225	0.145	0.126	0.216
Bulgaria	0.358	0.499	0.488	0.284	0.230	0.109	0.311	0.325
Canada	0.285	0.095	0.171	0.158	0.389	0.183	0.128	0.201
Chile	0.484	0.320	0.248	0.181	0.313	0.585	0.358	0.355
China	0.495	0.339	0.325	0.518	0.629	0.380	0.516	0.458
Taiwan	0.181	0.301	0.189	0.125	0.966	0.990	0.333	0.441
Croatia	0.293	0.168	0.113	0.391	0.478	0.110	0.248	0.258
Czech Republic	0.136	0.184	0.155	0.103	0.121	0.161	0.073	0.134
Denmark	0.288	0.313	0.294	0.208	0.209	0.311	0.220	0.264
Finland	0.220	0.171	0.200	0.188	0.266	0.118	0.133	0.185
France	0.184	0.113	0.143	0.124	0.279	0.314	0.181	0.191
Germany	0.249	0.230	0.204	0.175	0.240	0.264	0.189	0.221
Iceland	0.259	0.346	0.368	0.423	0.576	0.891	0.380	0.464
India	0.364	0.729	0.661	0.621	0.555	0.431	0.485	0.550
Ireland	0.175	0.130	0.123	0.138	0.486	0.251	0.133	0.205
Israel	0.196	0.220	0.248	0.275	0.243	0.211	0.228	0.231
Japan	0.136	0.253	0.198	0.246	0.879	0.700	0.186	0.371
South Korea	0.745	0.410	0.391	0.599	0.651	0.838	0.590	0.604
Latvia	0.350	0.255	0.236	0.299	0.259	0.334	0.199	0.276
Lithuania	0.170	0.173	0.149	0.166	0.188	0.063	0.275	0.169
Mexico	0.851	0.770	0.760	0.721	0.913	0.703	0.355	0.725
Norway	0.150	0.215	0.246	0.251	0.253	0.308	0.269	0.241
Philippines	0.661	0.754	0.923	0.740	0.846	0.439	0.536	0.700
Poland	0.105	0.078	0.104	0.291	0.379	0.583	0.266	0.258
Russia	0.313	0.124	0.130	0.245	0.300	0.186	0.079	0.198
Slovakia	0.220	0.154	0.158	0.166	0.159	0.284	0.083	0.175
Slovenia	0.225	0.268	0.190	0.246	0.265	0.738	0.240	0.310
South Africa	0.819	0.334	0.309	0.358	0.204	0.579	0.169	0.396
Sweden	0.119	0.169	0.175	0.171	0.400	0.124	0.180	0.191
Switzerland	0.268	0.254	0.236	0.171	0.405	0.604	0.269	0.315
Turkey	0.730	0.583	0.521	0.151	0.153	0.466	0.501	0.444
Great Britain	0.236	0.278	0.300	0.175	0.145	0.195	0.186	0.216
United States	0.120	0.143	0.180	0.320	0.246	0.399	0.175	0.226
Venezuela	0.574	0.316	0.174	0.499	0.498	0.626	0.483	0.453
Average	0.325	0.287	0.275	0.288	0.385	0.391	0.264	0.317

7. COMPUTATIONAL ISSUES IN COMPUTING THE DDI

The original work of Blasius and Thiessen (2012) on the DDI relied on the default values for terminating the iterations in the CatPCA procedure in SPSS Categories (version 22) which are comparable to those in homals (de Leeuw and Mair, 2009), both on a termination threshold value of 1e-05. homals is a package written in R which is a freely available statistical software environment (R Core Team, 2016). Within homals, we started to implement an algorithm for computing the DDI as well as simulating "good" and "bad" questionnaire data based on a latent-variable model (Nenadić and Blasius, in preparation).

As a preliminary result, we discovered that the obtained quantifications are sensitive to the choice of the threshold value for terminating the iterations. Choosing a value below 1e-09 can change the results quite drastically if the distribution of the margins is strongly skewed, only below this value, the quantifications appear to be stable under various distributions of the input data. Therefore, the upper bound of the DDI remains subject to further investigation.

8. CONCLUSION

For measuring the "quality of ordinal data", we further developed the Dirty Data Index (DDI) as first published by Blasius and Thiessen (2012). The DDI is based on the differences between the quantification areas from CatPCA (from the standard normal distribution) and the empirical cumulative frequencies (midpoint areas). The index is standardized between 0 and 1, whereby high values indicate poor data. Random data provide values – depending on the underlying distribution (u-shaped, normal, …) – between 0.5 and 0.7.

Applying the DDI to the ISSP 2012 (family and changing gender items), we could show that there are large differences in the quality of data between the countries and between the questions. The substantive solutions are close to previous solutions reported for the same countries, but different data (Blasius and Thiessen 2006, 2012). However, as recently discovered, the CatPCA quantifications might be unstable when using the default values for termination the iterative procedure when the data are heavily skewed. Therefore, we recommend to check the quantification values for strong outliers that can be caused by a very few cases in single categories at the beginning or the end of the scale. In general, it might be best to avoid such small numbers by combining the respective categories.

REFERENCES

Biemer, P.P. and Lyberg, L. (2003). *Introduction to survey quality*. Hoboken, NJ: Wiley.

Blasius, J. and Gower, J.C. (2005). Multivariate prediction with nonlinear principal components analysis: Application. In *Quality and Quantity. 39*: 373–390.

Blasius, J. and Thiessen, V. (2001). Methodological artifacts in measures of political efficacy and trust: A multiple correspondence analysis. In *Political Analysis*. *9*: 1-20.

Blasius, J. and Thiessen, V. (2006). A three-step approach to assessing the behavior of survey items in cross-national research. In M. Greenacre and J. Blasius, editors, *Correspondence Analysis and Related Methods*. Boca Raton, Florida: Chapman & Hall: 433-453.

Blasius, J. and Thiessen, V. (2009). Facts and artifacts in cross-national research: The case of political efficacy and trust. In M. Haller, R. Jowell and T. Smith, editors, *Charting the Globe. The International Social Survey Programme 1984-2009*: 147-169.

Blasius, J. and Thiessen, V. (2012). *Assessing the Quality of Survey Data*. London: Sage.

Blasius, J. and Thiessen, V. (2013). Detecting poorly conducted interviews. In P. Winker, N. Menold and R. Porst, editors, *Interviewers' Deviations in Surveys: Impact, Reasons, Detection and Prevention*. Frankfurt: Peter Lang: 67-88.

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

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

De Leeuw, J. (2006). Nonlinear principal component analysis and related techniques. In M. Greenacre and J. Blasius, editors, *Multiple Correspondence Analysis and Related Methods*. Milton Park, UK: Chapman and Hall/CRC: 107-133.

De Leeuw, J. and Mair, P. (2009). Gifi methods for optimal scaling in R: The package homals. In *Journal of Statistical Software. 31*(4): 1-20. URL http://www.jstatsoft.org/v31/i04/.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. New York: Wiley.

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

Krosnick, J.A. (1999). Survey research. In *Annual Review of Psychology. 50*: 337–367.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Saris, W. and Gallhofer, I.M. (2014). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Hoboken, New Yersey: Wiley.

Scheuch, E.K. (2000). The use of ISSP for comparative research. In *ZUMA-Nachrichten*. *47*, 64–74.

Simon, H.A. (1957). *Models of Man*. New York: Wiley.

Thiessen, V. and Blasius, J. (2008). Mathematics achievement and mathematics learning strategies: Cognitive competencies and construct differentiation. In *International Journal of Educational Research. 47*: 362-371.

Thiessen, V. and Blasius, J. (2016). Another look at survey data quality. In C. Wolf, D. Joye, T.W. Smith and Y. Fu, editors, *Sage Handbook of Survey Methodology*. Los Angeles: Sage: 613-629.

West, B.T. and Blom, A.G. (2016). Explaining Interviewer Effects: A Research Synthesis. In *Journal of Survey Statistics and Methodology (Online First)*: 1-37, Doi.org/10.1093/jssam/smw024.