Detecting Fabricated Interview Data in a Two-Level Approach Using the Mahalanobis Distance

Article · September 2018

CITATIONS
0

READS
41

1 author:

Robert Kuchen
Universitätsmedizin der Johannes Gutenberg-Universität Mainz
3 PUBLICATIONS **1** CITATION
SEE PROFILE

Detecting Fabricated Interview Data in a Two-Level Approach Using the Mahalanobis Distance

Robert Kuchen

2018-09-27

1 Introduction

Surveys rely on accurate interview data. Depending on the study, to obtain statistically trustworthy results, dozens or hundreds of interviews need to be conducted, which can normally only be achieved by hiring a large number of (usually freelancer) interviewers. In most cases, interviewers are payed for each completed interview, are not involved in the study and do not benefit from correct study results (Blasius & Friedrichs, 2012). They are rather interested in maximizing their payment and minimizing their effort. This naturally generates the incentive to fabricate parts of or even whole interviews.

Schräpler and Wagner (2005) have shown that even a small proportion of such fabricated interviews has the potential to severely bias multivariate statistics. Naturally, statisticians have thus tried to come up with methods to identify fabricated interviews and/or dishonest interviewers, respectively. Even though these methods differ in parts considerably from each other, the general idea is always that fabricated have the tendency to differ from “real” questionnaires regarding certain characteristics (Bredl et al., 2012). The problem is then to find such appropriate characteristics, which we call “indicator variables” (IV), and how to filter those interviews that have suspicious values with regard to these IVs.

In this paper, we will use R to check whether there are suspicious interviews within the ESS 2016 Belgium. We will do so by applying the Mahalanobis distance (MaDi), which can be used to detect multivariate outliers within a set of observations (Mahalanobis, 1936). Furthermore, we will, unlike previous publications, try to identify suspicious observations both on interviewer ID and on respondent ID level. Both concepts have, to the best of our knowledge, not been applied in the context of interview falsification. We will start by discussing the characteristics of the underlying survey, the ESS 2016 Belgium. We will then go on to introduce previous studies on the topic and explain how our approach differs from them. Subsequently, we will demonstrate which IVs we included in our analysis. We will then continue with processing these variables and explain the theory behind the MaDi. Subsequently, we will apply the concept to identify suspicious observations on interviewer and respondent level. Finally, we will go into limitations of our approach and give a prospect on further research that can be conducted.

1.1 ESS

The European Social Survey (ESS) 2016 was a cross-sectional survey aimed at identifying the attitude and behavior of the population of 23 European countries (European Social Survey 8 - 2016). The interviewees were randomly selected among the population of each country,

the sample size per country varying from 880 for small to 2852 for bigger countries. The questionnaire is comprised of eight modules that cover a wide range of topics. On the one hand, respondents were asked to provide information regarding their circumstances, such as educational background, employment and marital status or their financial situation and overall well-being. On the other hand, they were asked for their opinion on numerous topics, such as politics, the media, climate change or migration. Altogether there were 258 questions for the respondent. While there were also some open questions, most questions had to be answered in a multiple-choice manner and among those, most were on a Likert rating scale (ESS 8 - 2016 Variables and Questions). What is more, the ESS is characterized by a meticulous documentation. In addition to the questionnaire that is aimed at the interviewee, the ESS involves the interviewer questionnaire, in which the interviewer is required to answer questions regarding the interviewing procedure itself, such as whether the respondent often seemed to have been reluctant to answer questions (ESS 8 - 2016 Source Questionnaire). In the contact form, the interviewer is required to answer numerous meta questions, such as how many times the interviewee had to be contacted before the interview could be conducted. Furthermore, the date of the interview and the interview duration are asked (ESS 8 - 2016 Data Protocol).

Since the ESS 2016 contains several country specific questions and there are also other regulations that differ from country to country, we decided to choose only one from the 23 participating countries: Belgium.

ESS Belgium: The ESS 2016 Belgium was conducted by the Kantar TNS, Brussels and supervised by three national coordinators. Altogether, 138 freelance interviewers (103 of which had already previously worked for the ESS) administered 1765 interviews (*we deleted one interview because of crucial missing values*), all of which were conducted in a computer assisted personal interview (CAPI) fashion. The number of interviews varies greatly among interviewers. To illustrate this, Table 1.1 contains the number of interviews the topmost 15 interviewers conducted.

Table 1.1: No. of interviews per interviewer

Interviewer ID	1001	1002	1003	1004	1005	1006	1007	1008	1009	1010	1011	1012	1013	1014	1015
No. Interviews	29	14	5	3	20	14	4	11	4	9	8	2	16	15	6

Interviewers were paid 60 Euro for each completed, valid interview, 5 Euro for every completed and validated contact form and a bonus arrangement of 6 Euro per completed, valid interview conducted by interviewers who achieved an overall response rate above a certain threshold, have been assigned at least 18 addresses and meet certain interview quality standards and the requirements on contact registration and timings (ESS 8 - 2016 Documentation Report). We thus see that there is a general incentive to fabricate interviews, since not only are interviewers paid per completed interview, but they might even receive a bonus payment if they complete a certain number of interviews and have a relatively high response rate.

Lastly, it should be said that 437 interview-back-checks were conducted by telephone, of which 436 could confirm the interview results (ESS 8 - 2016 Documentation Report). Thus, many of the interviews were already proven to have been rightfully conducted. At the same time, this means that there are still more than 1300 interviews left that could have been fabricated.

1.2 At Risk Interviewer

Our goal is to identify possibly fabricated interview data. Yet, it should be clear that the only method to become certain that an interview was rightly conducted would be to reinterview the corresponding person. To reinterview every single interviewee is, however, too costly and usually only a small proportion of all interviewees can be contacted again or reinterviewed (Forsman & Schreiner, 1991).

We do thus not aim at detecting falsifiers with a 100% certainty, but rather at identifying interviews or interviewers that/who seem more suspicious than others and would be the preferred target for reinterviewing/interview-back-checks. Following Hood and Bushery (1997), we call these interviewers *at risk interviewers* (ARI).

1.3 Previous Studies and our Approach

There are some scientific papers that already dealt with identifying fabricated interviews.

Menold & Kemper (2014) used logistics regression to identify the IVs that have the strongest potential to separate fabricated from real interviews. They tested a high number of IVs, which they classified into meta-data, formal and content related indicators. Obviously, to apply logistic regression one requires a dependent binary variable; so, in this case a variable that indicates whether a certain interview was fabricated or not. Thus, Menold & Kemper instructed several “interviewers” to forge interviews. They then regressed this binary variable on the numerous indicator variables and compared their statistical significance. Even though this supervised approach is different from the purpose of this paper, it proves highly useful in providing possible IVs.

Closer to our approach is the study conducted by Bredl et al. (2012). While they also knew in advance which interviewers had fabricated their interviews, they applied an unsupervised approach assigning each interviewer to one of two clusters. One was supposed to contain unsuspicious interviewers, the other one ARIs, who could in theory be the preferred targets for re-interviewing. They used four different IVs, three of which were also included by Menold & Kemper. Bredl et al. were able to classify most interviewers correctly. Even though the approach worked in the case of their 13 interviewers, we consider this approach to be problematic. Cluster analysis has the goal to partition all observations (in this case interviewers) in a way that the observations within one cluster are as similar as possible to each other, while the observations of different clusters should be as different as possible. Thus, Bredl at el. assume that falsifiers form a homogeneous group of interviewers. However, this does not need to be the case. It is very well possible that one falsifier might stand out regarding some IV values, while another falsifier could be detected by some other IV. We did therefore not apply clustering.

Schäfer et al. (2004) used only two IVs in their paper and calculated for all interviewers the p-values. The only shortcoming of this approach is the fact that they did not combine the information that both IVs contained. It is very well possible for a multivariate observation to stand out if all variable values are considered together, while being unsuspicious on single variable level.

Our goal too is to detect interviews or interviewers that are more suspicious than others. If we assume most interviews were truthfully conducted (which should be given, if the number of interviews is large enough), being more suspicious means that we are looking for interviews that have an unusual combination of responses and conspicuously differ from most other interviews. Statistically speaking, we are thus looking for multivariate outliers. For this reason, we decided to use the Mahalanobis distance, which is a simple yet powerful method to detect those.

What is more, except Menold & Kemper all papers only tried to detect suspicious data on interviewer ID level by averaging the IV values of all interviews that one interviewer has conducted. However, it is believed that most falsifiers do not fabricate all but only some of their interviews (Schäfer et al., 2004). Usually they will conduct some real interviews and based on that experience they will fabricate the remaining interviews. The other approach would be to detect suspicious respondent/single interviews. However, having several unusual among dozens or hundreds of interviews is to be expected and nothing unusual. It only begins to become suspicious, if many of those unusual interviews were conducted by the same interviewer.

Our approach is therefore to avoid the shortcoming of both methods and to detect outliers both on interviewer and respondent level. We will work with one dataset on respondent level, which is comprised of 1765 observations (one for each respondent), and a second dataset, which only contains 138 observations, one for each interviewer (we will also average over all interviews conducted by the same interviewer), using the information of both datasets

2 Variable Selection, Summary and Processing

2.1 Used Indicator Variables

To identify useful IVs, it is helpful to understand the mindset of falsifiers. On the one hand, they want to minimize their time and cognitive effort. On the other hand, they want to avoid being detected. Hence, they aim at filling in the questionnaires as fast as possible, without raising suspicion. This will play a role in most IVs we will introduce in the following.

In this section, we will introduce the nine IVs we use to determine the MaDi of each interviewer. We first explain what each variable measures, before describing the motivation behind including each IV and discussing how we expect the IV values of a falsifier to differ from those of honest interviewers. Subsequently, we explain in detail how the variable values of a certain IV are obtained. Lastly, we summarize the distribution of the values that were obtained for each variable, one time on respondent and one time on interviewer level.

2.1.1 Extreme Response Style

What is being measured: Extreme Response Style (*ERS*) measures the number of extreme responses in the case of answers on a rating scale (most of which are Likert scales). For instance, in the case of a 1 to 5 scale, 1 and 5 would be considered extreme.

Rationale and expectation: Falsifiers want to deliver interviews that are as unsuspicious as possible. Apparently, many falsifiers believe that giving extreme answers might raise suspicion and thus tend to avoid these responses. Therefore, it is generally assumed that falsifiers tend to have fewer extreme responses in their questionnaires (Schäfer et al, 2004).

Detailed approach: This variable gives the proportion of extreme responses for variables whose answers involve a rating scale, often a Likert scale. Altogether, there were 163 such items. Table 2.1 shows the number of items that involve a certain number of response options.

Table 2.1: No. of questions with a certain no. of response options

No. of Response Options	4	5	6	7	10	11
No. of Questions	26	49	35	5	1	47

While in the case of four response options (1-4 scale) 50% of the answers are considered extreme answers, only 2/11 are done so in the case of eleven response options. To account for this, we weighted all items according to their response options scale, assigning questions with eleven response options a weight of 1 and questions with four responses options, for instance, a weight of 4/11. If all 163 rating scale items were answered with extreme response options, this would thus result in an indicator score of $92.\overline{63}$. *ERS* gives the proportion of points that were reached compared to the maximal possible score.

Data Summary: Tables 2.2, Table 2.3 and Figure 2.1 sum up the values obtained by the $n = 1765$ interviews. The vertical dashed brown line indicates the sample mean. The two dotted/dashed vertical orange lines represent the lower and upper outlier threshold based on the formula:

$$T_{lower} = Q_1 - 1.5*IQR,\;\;\;\;\;\;\;\;\; T_{upper} = Q_3 + 1.5*IQR \tag{2.1} \label{eq:2.1}$$
where Q_1 and Q_3 are the first and third quantile respectively and IQR is the interquantile range (*the dashed lines are only visible, when they are in the range of the observed IV values*).

Table 2.2: ERS summary output, Respondent level

ERS	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.022	0.101	0.137	0.158	0.195	0.606

Table 2.3: ERS summary output, Interviewer level

ERS	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.069	0.138	0.153	0.159	0.183	0.269

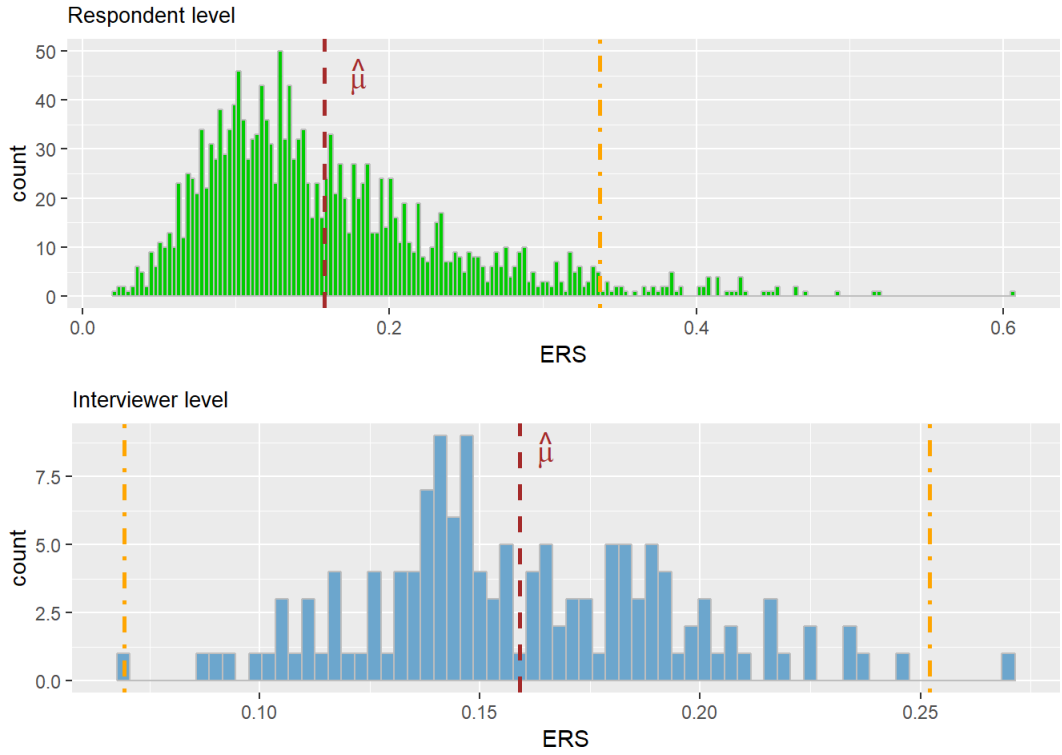

Figure 2.1: Histograms ERS

It can be observed that the distribution on respondent level has a much higher variance. This is no surprise, considering that the IV values on interviewer level are obtained by averaging over all interviews that were conducted by the same interviewer.

2.1.2 Middle Response Style

What is being measured: Middle Response Style (*MRS*) measures the proportion of middle responses in the case of answers on a rating scale that have an uneven number of response alternatives. For instance, in the case of a 1 to 5 scale, 3 would be considered a middle response.

Rationale and expectation: It is believed that falsifiers tend to avoid extreme responses. With the same rationale, it is assumed that falsifiers will have a higher number of middle responses than honest interviewers (Menold & Kemper, 2014).

Detailed approach: We sum up the number of middle responses in the case of answers on a rating scale with an uneven number of response options. As can be observed in Table 2.1, there are 101 such items in the ESS questionnaire (questions with five, seven or eleven response options). As in the case of *ERS*, we weighted all items according to their number of response alternatives, so questions with seven response options, for instance, were only assigned a weight of 7/11. If all those 101 items were answered with a middle response, this would result in an indicator score of \(66.\overline{45}\). *MRS* gives the proportion of score points that were reached compared to that score.

Data Summary:

Table 2.4: MRS summary output, Respondent level

MRS	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.039	0.13	0.169	0.176	0.212	0.668

Table 2.5: MRS summary output, Interviewer level

MRS	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.124	0.16	0.171	0.177	0.189	0.545

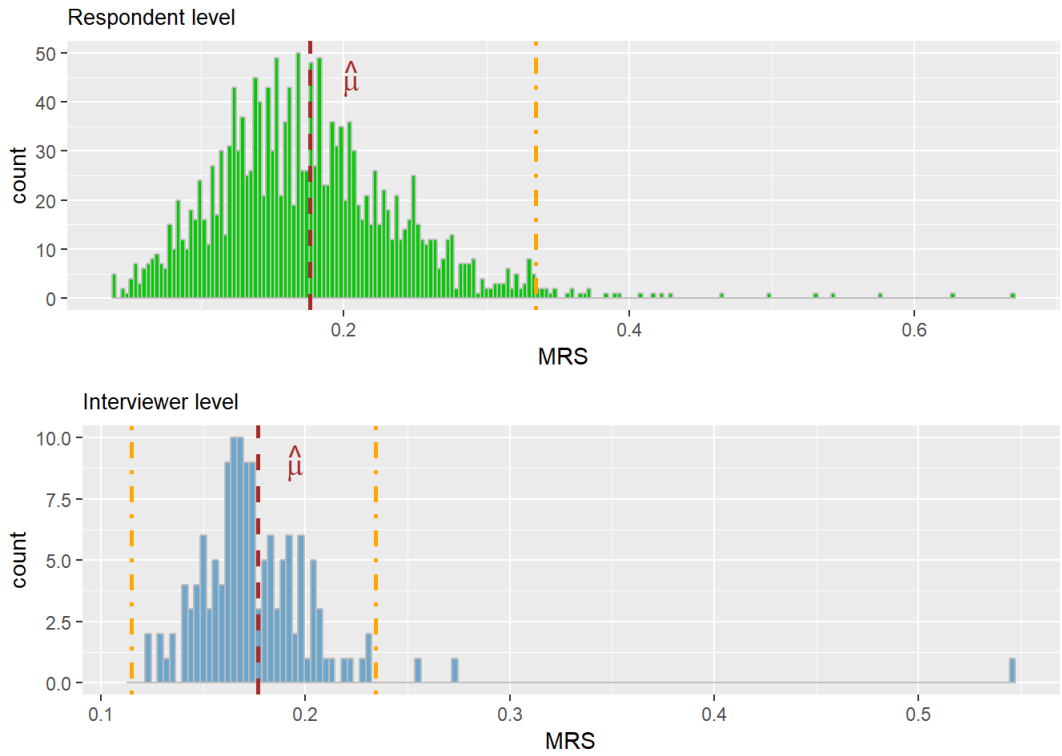

Figure 2.2: Histograms MRS

2.1.3 Variance

What is being measured: Variance (VAR) measures the variance of all answers to rating scale questions.

Rationale and expectation: Since falsifiers are believed to avoid extreme responses and give more middle responses than honest interviewers, they are assumed to have a lower variance in their responses to rating scale questions (Schäfer et al, 2004).

Detailed approach: We used all 163 rating scale items from Table 2.1. In order to make the variance of items with a different number of response alternatives comparable, we standardized them, so that each rating scale variable has a mean of 0 and variance of 1 over all respondents. We then calculated the variance of all answers from one respondent (*The standardization is only needed for this IV. We continue with unstandardized values of all rating scale variables*).

Data Summary:

Table 2.6: VAR summary output, Respondent level

VAR	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.353	0.722	0.906	0.989	1.162	2.72

Table 2.7: VAR summary output, Interviewer level

VAR	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.493	0.906	0.991	0.998	1.101	1.434

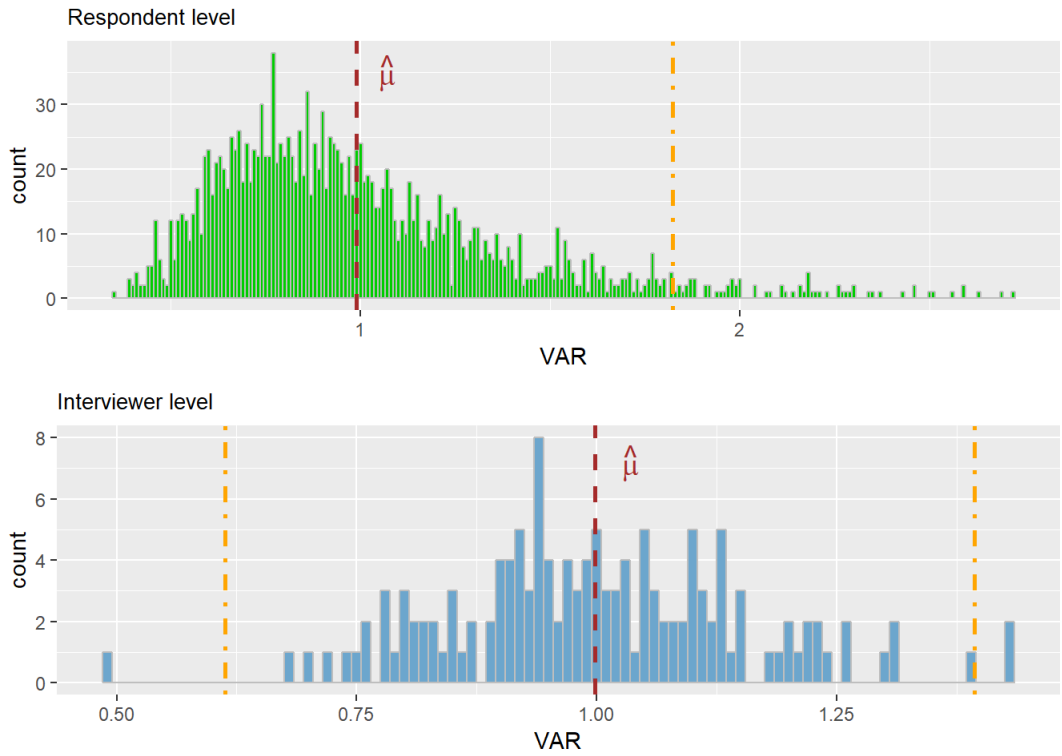

Figure 2.3: Histograms VAR

2.1.4 Minutes

What is being measured: The length in minutes an interview took is measured.

Rationale and expectation: This variable was proposed by Bushery, Reichert, Albright and Rossiter (1999) and Murphy et al. (2004). Falsifiers might not know how long a real interview may take. If they fill in a time that is much shorter or much longer than those filled in by other interviewers, this might be suspicious. Falsifiers might fill in a number close to 55 minutes, since this is the time an interview was supposed to take according to the creators of the questionnaire. However, the interviews took actually on average about 67 minutes (Table 2.8).

Detailed approach: We use the number of minutes the whole interview took, which had to be filled in by the interviewer. All interviewers also had to document how long each of the nine question modules took. We did not use this information (*to be precise, we tried to use it, but the obtained results did not differ significantly*).

Data Summary:

Table 2.8: Minutes summary output, Respondent level

Minutes	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	11	55	65	66.98	76	181

Table 2.9: Minutes summary output, Interviewer level

Minutes	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	47.96	61.42	66.86	68.16	73.78	98.75

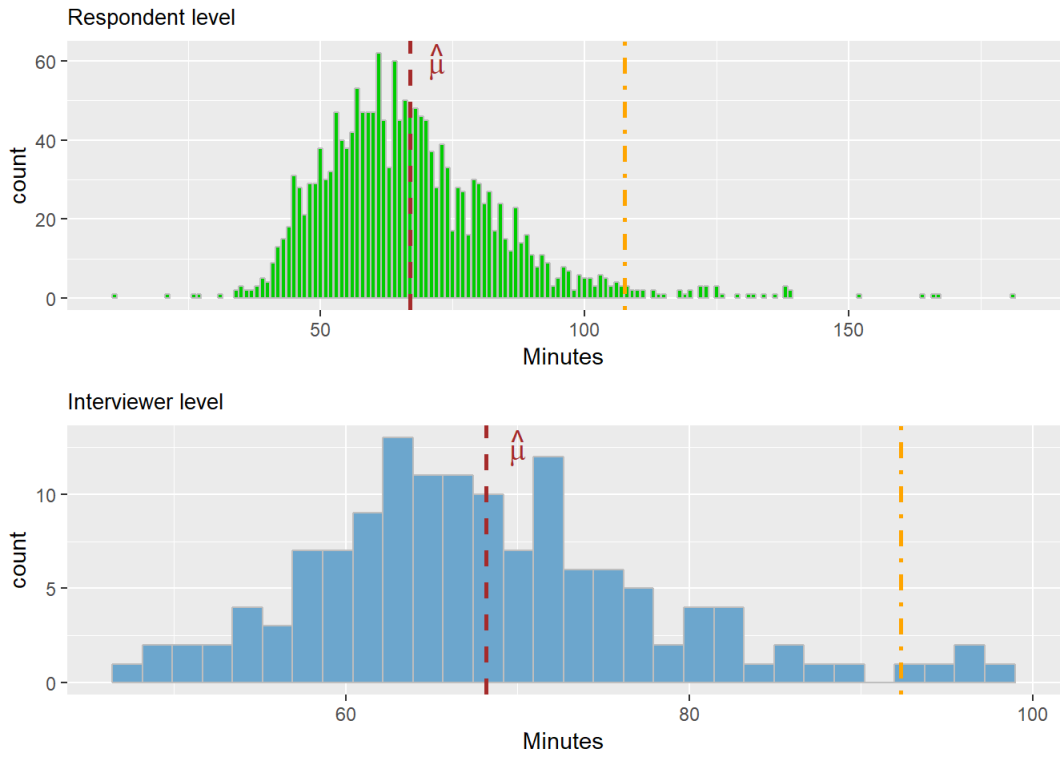

Figure 2.4: Histograms Minutes

2.1.5 Rounding

What is being measured: The number of times in which rounded answers were given to open-ended numerical questions is measured.

Rationale and expectation: Following Menold & Kemper (2014), it is expected that falsifiers do not round as often as real respondents, as they might suspect that too much rounding would seem suspicious.

Detailed approach: There were four items within the ESS that could be used to measure the rounding behavior within one interview. The first two questions asked for the time in hours and minutes the respondent spends following news on politics and using the internet. Following Menold & Kemper, we considered an answer to these questions to be rounded, if it was divisible by 30 (i.e. 30, 60, 90, 120 … minutes). In the other two questions, respondents were asked to give the total number of hours they (question 1) and their partner (question 2) worked per week in their main job, overtime included. In this case, we regarded an answer to be rounded, if it was divisible by 5. We did not include the questions on the hours spend per week excluding overtime, since we believed respondents were unlikely to round there, as they are likely to know their exact contract hours.

Data Summary:

Table 2.10: Rounding summary output, Respondent level

Rounding	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0	1	2	2.063	3	4

Table 2.11: Rounding summary output, Interviewer level

Rounding	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.333	1.879	2.08	2.076	2.286	3

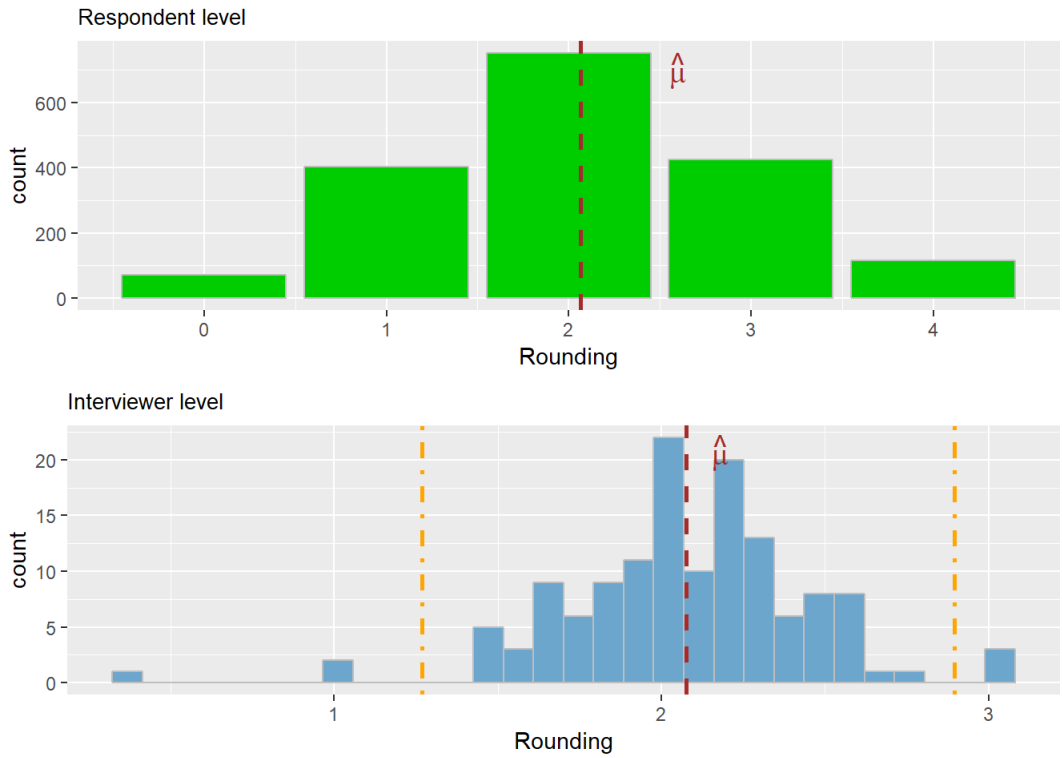

Figure 2.5: Histograms Rounding

2.1.6 Telephone

What is being measured: Telephone (TELE) indicates whether a respondent disclosed his or her telephone number.

Rationale and expectation: This IV was first used by Hood and Bushery (1997). The idea is that falsifiers are more reluctant to fill in a telephone number for each respondent, since they either have only partially conducted the interview and are afraid that this is detected when the respondent are phoned or they have fabricated the whole interview and do not know the telephone number. We therefore expect the proportion of absent telephone numbers to be higher for falsifiers.

Detailed approach: The ESS contact data form explicitly asks for the telephone number of each interviewee. We thus simply used this information for each interview.

Data Summary:: In Tables 2.12-2.14 and Figure 2.6, 0 (1) indicates that no (a) Telephone number was provided.

Table 2.12: TELE summary output, Respondent level

TELE	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0	1	1	0.937	1	1

Table 2.13: TELE summary output, Interviewer level

TELE	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.333	1	1	0.949	1	1

Table 2.14: TELE distribution, Respondent level

Value TELE, Respondent level	**0**	**1**
No. observed	112	1654

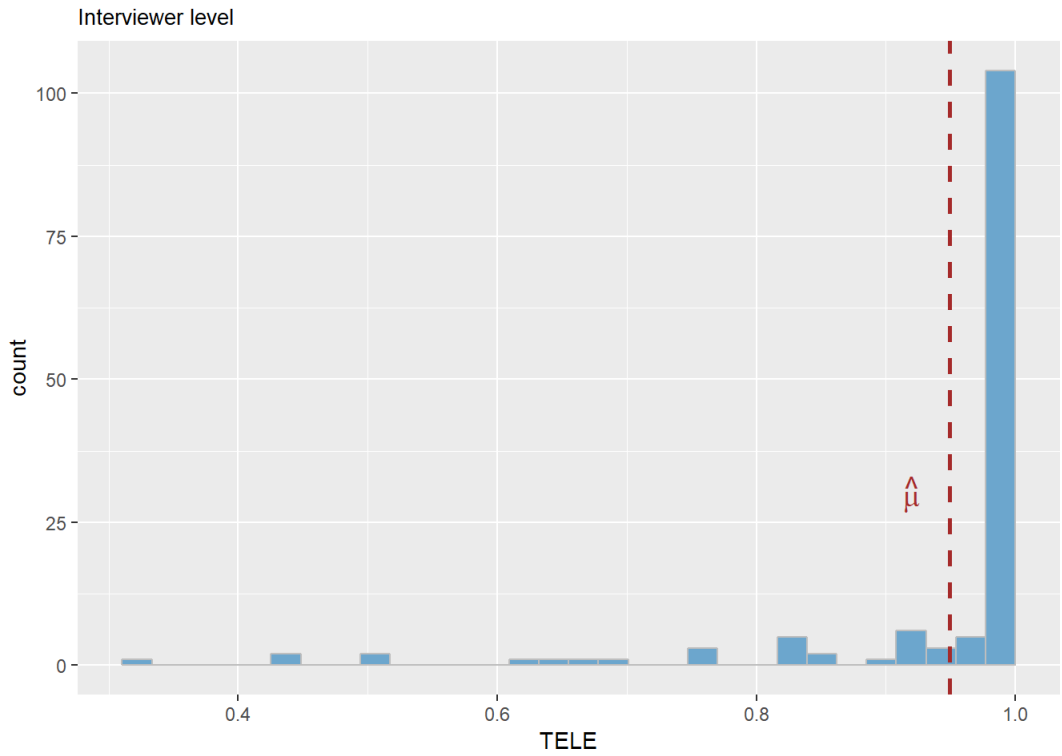

Figure 2.6: Histogram TELE

2.1.7 Filter

What is being measured: Filter questions are questions that can, depending on the given answer, change the subsequent pathway of the interview. If, for instance, the respondent answers that he or she is in a relationship, he or she might be asked to give information on his or her partner. If, however, the respondent answers that he or she is single, no further questions need to be answered. When there are several of these filter questions, there will be many possible different pathways through the interview. We measured the length of an interview's pathway compared to the longest possible path.

Rationale and expectation: Falsifiers are interested in minimizing their time and efforts to fill in the questionnaires. They might thus regard filter questions as a good opportunity to save time, by choosing pathways that are relatively short. Consequently, the longer the chosen pathway is, the less suspicious an interview becomes.

Detailed approach: We counted a number of 28 filter questions in the ESS. While some of those are single level, others are multilevel filter questions, i.e. to arrive at such a filter question some other filter question(s) needed to be answered accordingly. All in all, these filter questions result in a total of 1'783'627'776 possible pathways through the interview. The longest possible filter path through the interview is 111 questions longer than the shortest possible path. We thus counted what proportion of the longest possible path was taken in each interview.

2.1.7.1 Data Summary:

Table 2.15: Filter summary output, Respondent level

Filter	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.063	0.216	0.261	0.26	0.306	0.496

Table 2.16: Filter summary output, Interviewer level

Filter	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.198	0.244	0.261	0.262	0.28	0.342

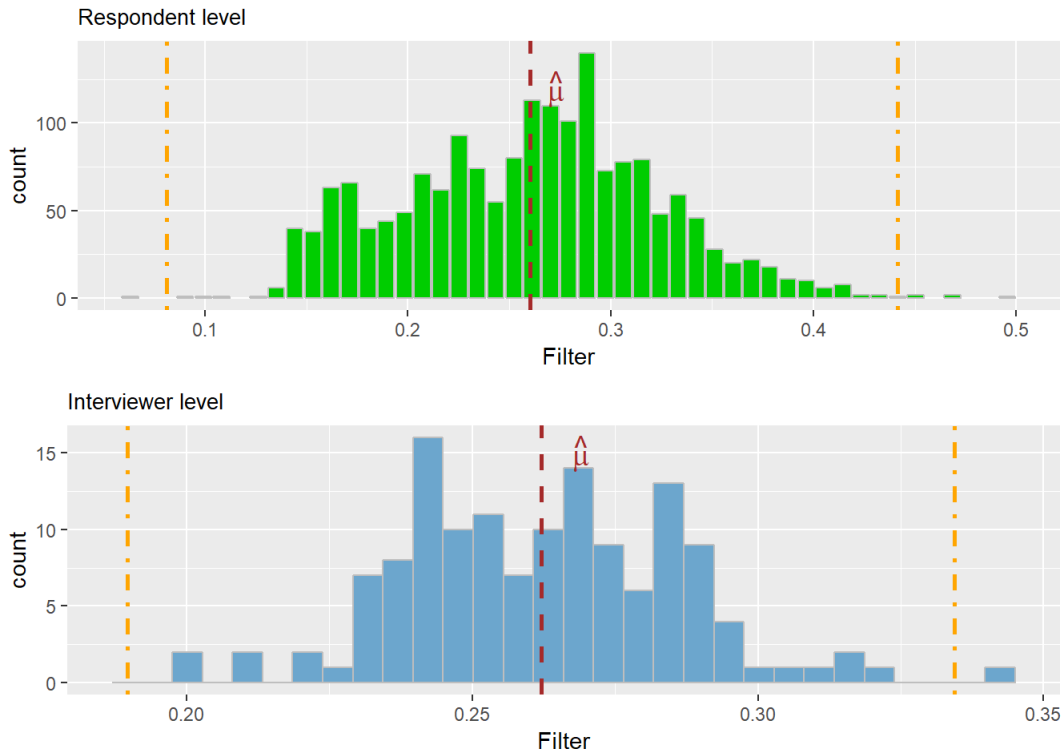

Figure 2.7: Histograms Filter

2.1.8 Open Questions Item non Response

What is being measured: Open Questions Item non Response (*OQ INR*) measures the number of missing values per respondent in the case of open-ended questions (questions that have no response alternatives given).

Rationale and expectation: Following Menold & Kemper (2014), we distinguish between non response in the case of open questions and in the case of all other questions. Generally it is believed that falsifiers have a lower proportion of missing values, since they want to avoid raising suspicion. One might, however, also argue that they actually produce more missing values to save time. This might particularly be the case for open questions.

Detailed approach: We counted the number of missing values in the 28 open questions. We excluded filter-follow-up-questions, since those were already considered in the IV *Filter*.

Data Summary: The high proportion of missing values here is owed to the fact that a high portion of the questions asks for the birthdate of the 2nd, 3rd, 4th, …, 9th household member.

Table 2.17: OQ INR summary output, Respondent level

OQ INR	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.143	0.464	0.536	0.518	0.571	0.857

Table 2.18: OQ INR summary output, Interviewer level

OQ INR	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.345	0.493	0.517	0.515	0.538	0.652

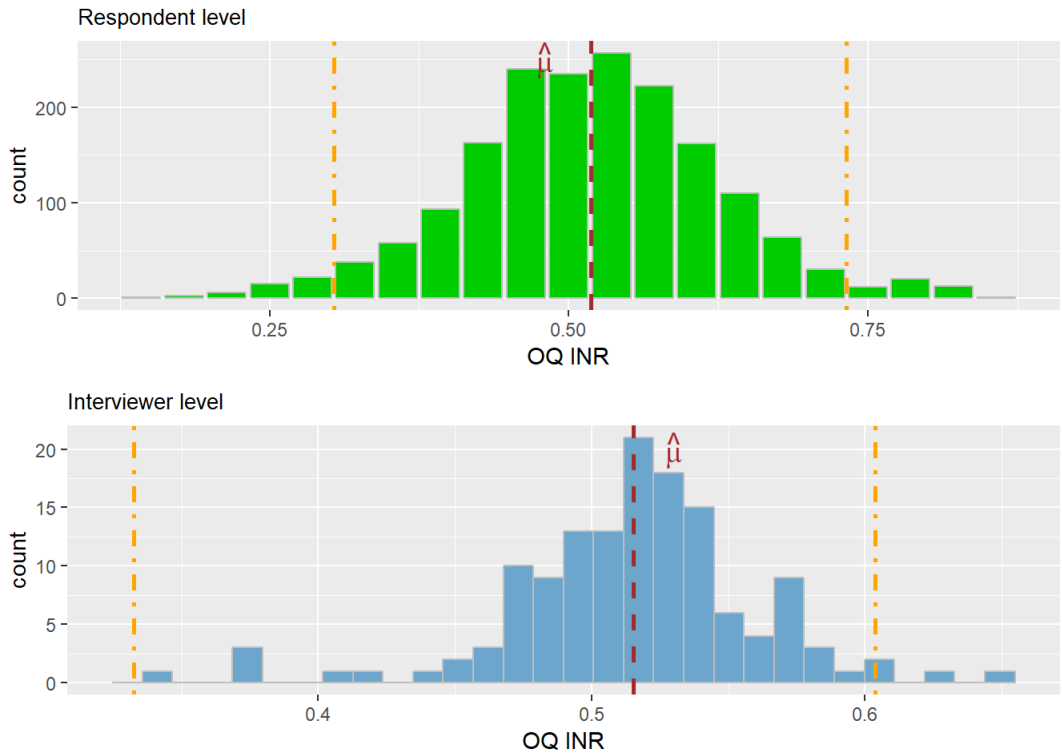

Figure 2.8: Histograms OQ INR

2.1.9 Other Questions Item non Response

What is being measured: Other Questions Item non Response (*Other INR*) measures the number of missing values per respondens for non-open-ended questions.

Rationale and expectation: We expect a lower proportion of missing values for falsifiers. This does, however, not necessarily be the case.

Detailed approach: We counted the number of missing values for the 223 open questions that were no at the same time filter-follow-up-questions. The variable indicates what proportion of the 223 questions is missing per interview.

Data Summary::

Table 2.19: Other INR summary output, Respondent level

Other INR	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.04	0.05	0.055	0.06	0.06	0.36

Table 2.20: Other INR summary output, Interviewer level

Other INR	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.048	0.055	0.058	0.06	0.061	0.109

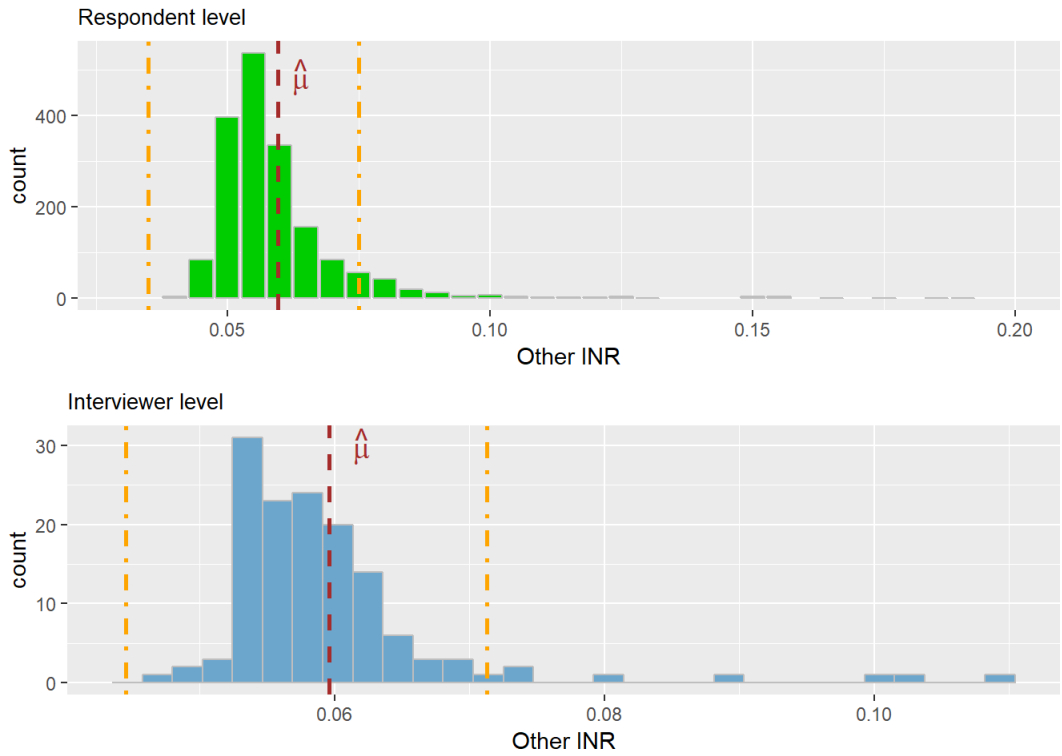

Figure 2.9: Histograms Other INR

2.1.10 Summary of Datasets

Tables 2.21 and 2.22 summarize the IV values of both levels. Table 2.21 shows five randomly sampled respondents and their IV values. Table 2.22 depicts the first five interviewers, whose values were obtained by averaging the IV values of all respondents assigned to the same interviewer.

Table 2.21: Subset of respondent level dataset, Five randomly chosen interviews

Resp_ID	Int_ID	Minutes	ERS	MRS	VAR	Rounding	Filter	OQ_INR	Other_INR	TELE
247	1027	68	0.165	0.168	0.960	3	0.288	0.429	0.050	1
1365	1119	70	0.384	0.183	1.356	1	0.207	0.607	0.055	1
1613	1133	68	0.285	0.109	1.798	3	0.288	0.536	0.060	1
2175	1091	72	0.044	0.326	0.448	3	0.315	0.464	0.055	1
2731	1035	74	0.089	0.075	0.620	3	0.270	0.464	0.055	1

Table 2.22: Subset of interviewer level dataset

Int_ID	Minutes	ERS	MRS	VAR	Rounding	Filter	OQ_INR	Other_INR	TELE
1001	70.7	0.100	0.164	0.763	1.759	0.256	0.532	0.057	1.000
1002	63.4	0.153	0.181	1.071	2.000	0.240	0.541	0.054	1.000
1003	52.2	0.237	0.146	1.121	2.200	0.270	0.493	0.058	1.000
1004	65.0	0.089	0.545	1.225	0.333	0.246	0.488	0.058	0.333
1005	72.6	0.193	0.143	1.104	2.000	0.232	0.548	0.058	1.000

Note that in many cases it is from a theoretical standpoint not entirely clear how exactly the IV values of a falsifier might differ from those of an honest interviewer (only in the case of the variable *Filter*, we believe that falsifiers are very unlikely to have higher values than honest interviewers). For instance, in the case of INR, one might expect falsifiers to have a higher proportion of missing values, since they want to minimize their effort. On the other hand, one cannot rule out that the tendency might be the complete opposite, so that falsifiers tend to have a lower proportion of missing values, believing that a high proportion might raise suspicion. The relationship is not only in theory not entirely clear, but also empirical findings are equivocal. For instance, in their first study Menold & Kemper (2014) found that against their expectations falsifiers had a higher proportion of INR, while in their second study they had a lower proportion of INR.

This is, however, for our approach immaterial. The MaDi merely quantifies to what degree a certain observation differs from the average observation of one population. We thus do not require any expectation on the response tendencies of falsifiers for all IVs.

2.2 Not Used Indicator Variables

There are many more IVs proposed in the literature. Here are some candidates that we did not use:

- Semi Open Questions: This was among others used by Menold & Kemper (2014) or Bredl at al.. Some questions have response alternatives that do not cover the range of possible answers. If the respondent wants to give such an answer, he or she must select the “Others, please specify” field and fill in the alternative. It is believed that falsifiers often avoid this response option, making this a powerful IV. However, there are no such questions in the ESS.
- Benford’s Law: The first digit of monetary number is shown to follow a special logarithmic distribution. Even though Swanson et al. (2003) showed that this law does not fully apply to surveys, since people tend to give rounded answer, it is still possible to compare the distribution of the monetary values filled in by a certain interviewer to the distributions of those filled in by other interviewers. The only monetary questions in the ESS had, however, to be answered by multiple choice, making the application of Benford’s law impossible.
- Primacy and Recency: In the case of closed-ended questions with nominal response alternatives, research has shown that people’s answers are influenced by the primacy and recency effect. In the case of visually presented vertical response options, people tend to prefer answers that are in the top of a list (primacy effect). In the case of an acoustical presentation, answers that were read out last are chosen more often. We did not include these IVs, since there were only a few cases for which we were not sure whether these effects would apply to. For instance, if an interviewee is asked to name the party he or she voted for in the latest election, it seemed more intuitive to us that he or she would not let his answer be influenced by the order of presentation. What is more, we were not sure to what extend these effects apply to ordered numerical answers, so we decided not to include these IVs at all.

2.3 Dimension Reduction

The three IVs *ERS*, *MRS* and *VAR* are based on very similar concepts, measuring the degree to which moderate were preferred over extreme answers. If we calculate the MaDi based on all three IVs, the results are likely to be biased towards that concept. We thus decided to apply principal component analysis (PCA), only using the first principal component in the analysis, so that we can compress as much information as possible in just one variable. The results of the PCA (performed with the standard R function *prcomp*) are shown in Table 2.23.

Table 2.23: First principal component analysis summary output

PCA NON DIF	PC1	PC2	PC3
Standard deviation	0.1596	0.04145	0.01854
Proportion of Variance	0.9251	0.06237	0.01248
Cumulative Proportion	0.9251	0.98752	1.00000

Since we conducted the PCA on three variables, there are up to three PCs. The first PC, which we call *NON DIF*(= non differentiation), explains about 92.5% of the variability in the three IVs, containing the better part of their information. Tables 2.24/2.25 and Figure 2.10 illustrate the distributions of *NON DIF* on both levels.

Table 2.24: NON DIF summary output, Respondent level

NON DIF	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	-0.646	-0.271	-0.086	0	0.179	1.769

Table 2.25: NON DIF summary output, Interviewer level

NON DIF	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	-0.448	-0.099	0.007	0	0.094	0.514

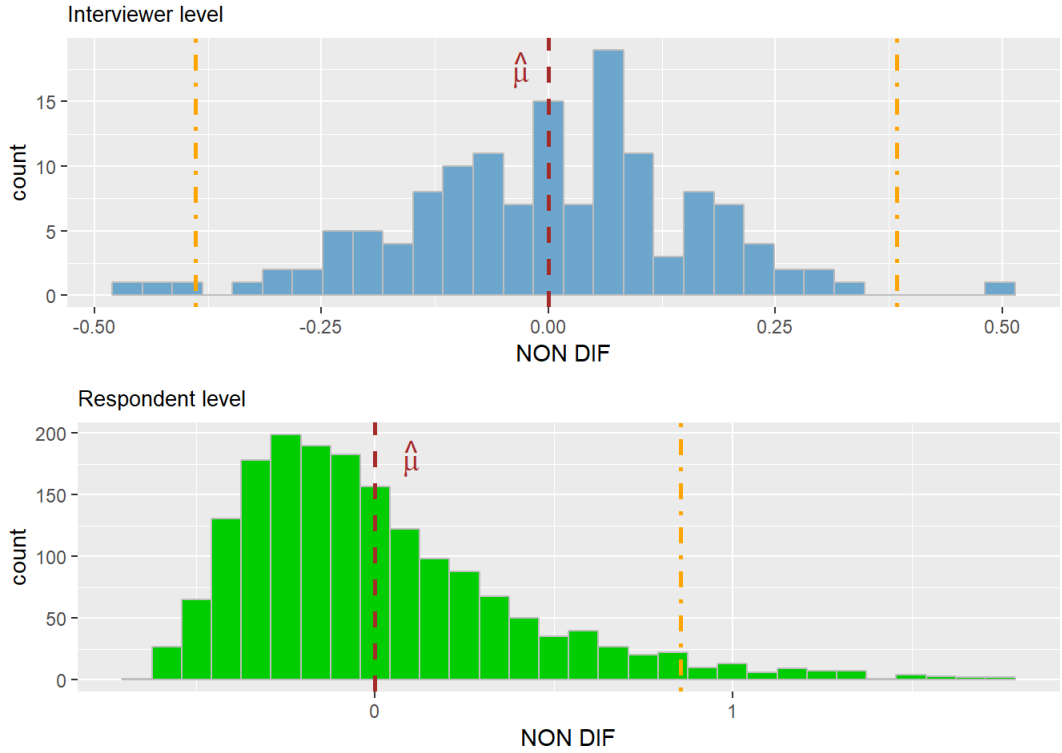

Figure 2.10: Histograms NON DIF

Similarly, we do not want the concept of item non response to unproportionally influence the Mahanalobis distance values. Consequently, we conduct a PCA with the two variables *OQ INR* and *Other INR*. The results are shown in Table 2.26.

Table 2.26: Second principal component analysis summary output

PCA INR	PC1	PC2
Standard deviation	0.0460	0.008657
Proportion of Variance	0.9658	0.034210
Cumulative Proportion	0.9658	1.000000

Again we only use the first PC, which we simply call *INR* and which explains almost all of the variance in the two IVs. Tables 2.27/2.28 and Figure 2.11 show the distributions of *INR* on both levels.

Table 2.27: INR summary output, Respondent level

INR	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	-0.375	-0.054	0.017	0	0.053	0.342

Table 2.28: INR summary output, Interviewer level

INR	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	-0.136	-0.022	-0.002	0	0.022	0.17

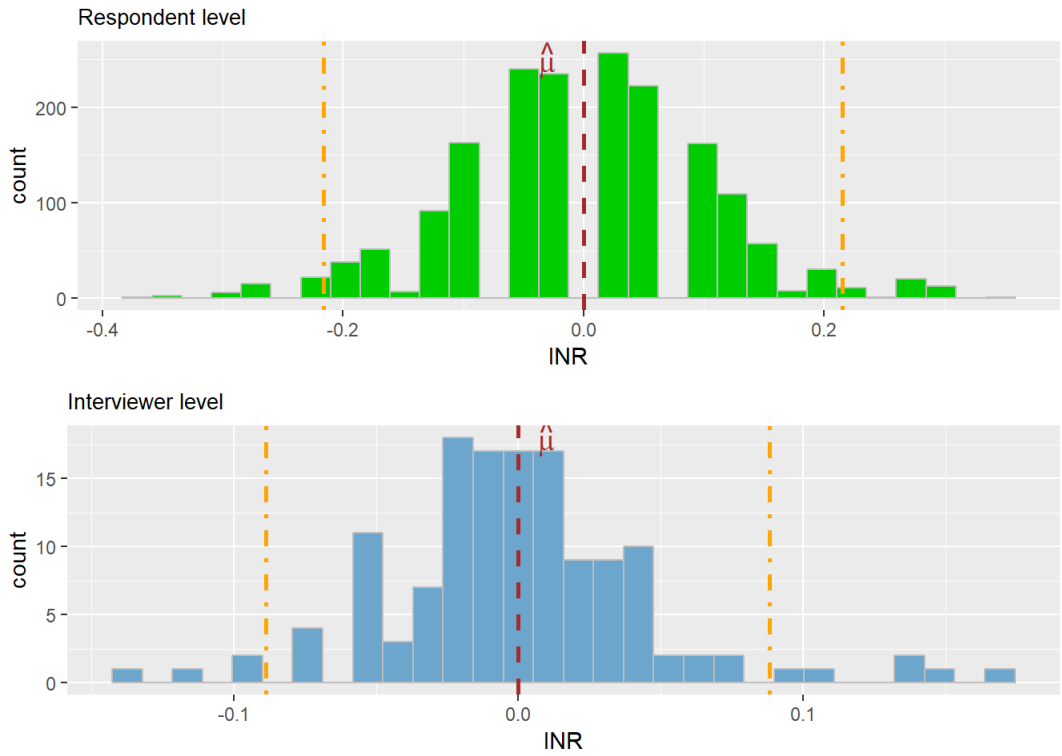

Figure 2.11: Histograms INR

3 Mahalanobis Distance

3.1 Theory

The main tool we use in this paper is the Mahalanobis distance (Mahalanobis, 1936), which is given by:

$$d_{M,i}(\mathbf{x_i},\mathbf{X}) = \sqrt{(\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol\Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})} \tag{3.1} \label{eq:3.1}$$
$$\mathbf{x}_i = (x_{1,i},x_{2,i},...,x_{p,i}) \quad E[\mathbf{X}] = \boldsymbol\mu = (\boldsymbol\mu_1,\boldsymbol\mu_2,...,\boldsymbol\mu_n)^\top \quad \mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2,..., \mathbf{x}_n)$$

where $\boldsymbol\mu$ is the mean vector and $\boldsymbol\Sigma$ the covariance matrix of a set of observations $\mathbf{X}$. It thus measures the distance of a certain observation $\mathbf{x}_i$ from the center of all observations $\boldsymbol{\mu}$. If this distance exceeds a certain threshold T_M, we can think of $\mathbf{x}_i$ as a multivariate outlier and, in our case, as a possibly suspicious observation. $\eqref{eq:3.1}$ includes the inverse of the covariance matrix $\boldsymbol\Sigma$, so that the contribution of a variable with a high variance to the distance d_{M} is proportionally weighted down. The MaDi is therefore unitless and scale-invariant and their is no need to standardize our IVs.

There is no general rule on which value to choose for the threshold T_M. There is, however, one useful relationship: If the set of observations $\mathbf{X}$ approximately follows a multivariate normal distribution, $d_{M,i}^2$ will be approximately χ^2-distributed with p degrees of freedom, where p equals the number of variables in $\mathbf{X}$ (Warren et al., 2011). In our case, there are two variables that are clearly not normally distributed: Rounding and TELE (Compare Figures 2.5 and 2.6). On that account, the six variables are impossible to follow a common multivariate normal distribution. However, Warren et al. (2011) showed that even when this condition is not fulfilled, the squared MaDi relatively robustly follows a χ^2-distribution. Therefore, we assume that the squared MaDi follows a χ_6^2-distribution (since six IVs remain after having conducted the PCA). Under this condition, we can use the 90%, 95% and 99% quantiles of the χ_6^2-distribution as threshold values to identify outliers. Table 3.1 contains these values:

Table 3.1: χ^2_6 - quantile values

χ^2_6 - Quantile	Value
90%	10.645
95%	12.592
99%	16.812

If the MaDi of a certain observation exceeds such a threshold value, it can be marked an outlier with an error probability of 10%, 5% or 1%.

3.2 Application

Having processed the data and explained the theory behind the MaDi, we are now in a position to calculate the MaDi of each interviewer and each respondent/interview, respectively. To perform the calculations, we use the default R-function *mahalanobis*.

3.2.1 Interviewer Level

Table 3.2 shows the summary output of all squared MaDis on interviewer level, Table 3.3 lists the number of interviewers whose MaDi values were below or above a certain significance level and Table 3.4 lists the 15 interviewers with the highest squared MaDi values (*note: even though the values were calculated based on the variables Minutes, Rounding, Filter, TELE, NON DIF and INR, we listed them together with the original variables to make interpretation easier*):

Table 3.2: Squared Mahalanobis distances summary output, Interviewer level

\(d_M\)	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.3191	2.4545	4.2961	5.9565	7.2196	42.1395

Table 3.3: No. of outliers per significance level, Interviewer level

Significance level	1%	5%	10%	Below 10%
No. of Interviewers	8	5	7	118

Table 3.4: 15 highest squared Mahalanobis distances, Interviewer level

Int_ID	Minutes	ERS	MRS	VAR	Rounding	Filter	OQ_INR	Other_INR	TELE	Mah_Dis
1004	65.0	0.089	0.545	1.225	0.333	0.246	0.488	0.058	0.333	42.14***
1058	96.0	0.164	0.165	1.029	2.000	0.321	0.345	0.050	0.667	26.77***
1038	66.6	0.226	0.198	1.232	2.286	0.208	0.520	0.059	0.429	25.51***
1099	59.0	0.126	0.230	1.104	2.500	0.311	0.464	0.102	0.500	23.69***
1052	66.5	0.217	0.133	1.151	2.500	0.203	0.536	0.062	0.500	22.86***
1132	73.1	0.161	0.205	0.991	1.632	0.246	0.523	0.061	0.447	17.23***
1049	72.2	0.269	0.142	1.434	1.500	0.230	0.652	0.056	1.000	17.22***
1083	66.0	0.146	0.167	0.697	3.000	0.198	0.536	0.060	1.000	17.06***
1128	78.0	0.178	0.140	0.898	3.000	0.275	0.375	0.048	1.000	15.91**
1007	94.8	0.151	0.148	1.048	1.500	0.212	0.607	0.061	1.000	13.5**
1073	74.7	0.069	0.166	0.493	2.000	0.261	0.488	0.055	1.000	13.43**
1139	76.0	0.105	0.183	0.834	2.000	0.258	0.631	0.055	1.000	13.28**
1015	68.0	0.187	0.232	1.264	2.333	0.318	0.375	0.058	1.000	13.23**
1050	96.7	0.208	0.147	1.311	1.571	0.283	0.508	0.052	1.000	12.38*
1055	77.0	0.165	0.124	1.034	2.400	0.288	0.371	0.050	1.000	12*

As can be seen in Table 3.3, there are 20 interviewers that are an outlier at least on the 10% level. This is a bit more than one would expect among 138 observations. Yet, a \(\chi_6^2\) distribution still provides a good approximation to the distribution of all squared MaDis (Figure 3.1).

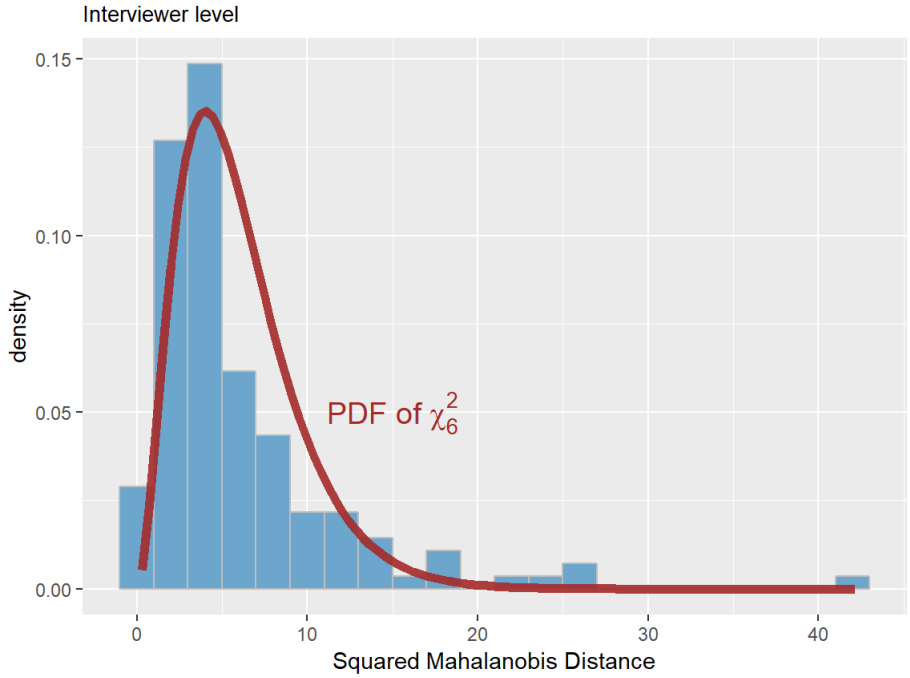

Figure 3.1: Histogram, Mahalanobis Distance Interviewer level

To provide some intuition on why some of the interviewers are regarded outliers, we will explain this by means of the examples of the two interviewers with the highest scores:

- Interviewer 1004: This male interviewer has by far the highest MaDi. This is mainly the case because he has marked the middle response alternative extremely often (*MRS* = 0.545, \(T_{upper}\) is 0.234), rarely provided rounded answers (*Rounding* = 0.33, \(T_{lower}\) = 1.268) and did only once provide the interviewee's telephone number. Yet, he submitted only three questionnaires.
- Interviewer 1058: This male interviewer has unusually high values for *Minutes* (though no univariate outlier according to \(\eqref{eq:2.1}\)) and for *Filter*. Since we believe, however, that falsifiers should have small values for *Filter*, the interviewer should not be considered suspicious, despite his high MaDi score. This example shows that one should not blindly trust the MaDi values, but must also apply some common-sense thinking. This interviewer too, has only submitted three questionnaires.

If we look at Table 3.5, we can see that most interviewers with high MaDI scores have only submitted very few questionnaires.

Table 3.5: No. of interviews per interviewer

Interviewer ID	1004	1058	1038	1099	1052	1132	1049	1083	1128	1007	1073	1139	1015	1050	1055
No. Interviews	3	3	7	2	2	38	4	1	2	4	3	3	6	14	5

This is problematic. We would be more certain to regard an interviewer as suspicious who has a high MaDi score and has submitted a considerable number of interviews. If an interviewer has conducted only a small number of interviews, his averaged IV values are naturally expected to have a higher variance. This is one of the reasons we believe the respondent level should be considered, too.

3.2.2 Respondent level

Tables 3.6, 3.7 and 3.8 can be interpreted equivalently to Tables 3.2, 3.3 and 3.4, this time on respondent level.

Table 3.6: Squared Mahalanobis distances summary output, Respondent level

\(d_M\)	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.188	2.387	4.223	5.997	6.721	58.448

Table 3.7: No. of outliers per significance level, Respondent level

Significance level	1%	5%	10%	Below 10%
No. of Interviewers	147	46	37	1535

Table 3.8: 15 highest squared Mahalanobis distances, Respondent level

Resp_ID	Int_ID	Minutes	ERS	MRS	VAR	Rounding	Filter	OQ_INR	Other_INR	TELE	Mah_Dis
1813	1050	181	0.373	0.141	1.882	0	0.396	0.536	0.050	1	58.45***

Resp_ID	Int_ID	Minutes	ERS	MRS	VAR	Rounding	Filter	OQ_INR	Other_INR	TELE	Mah_Dis
756	1091	166	0.407	0.120	2.498	2	0.225	0.607	0.075	1	43.26***
2019	1121	167	0.318	0.235	1.838	3	0.216	0.500	0.090	1	38.68***
1835	1109	123	0.385	0.090	2.113	1	0.162	0.821	0.065	0	37.51***
3036	1092	80	0.179	0.178	1.182	0	0.468	0.250	0.065	0	37.07***
2497	1119	164	0.142	0.257	1.306	0	0.171	0.643	0.055	1	36.09***
2896	1019	139	0.182	0.263	1.003	1	0.270	0.607	0.055	0	34.7***
680	1078	34	0.088	0.151	1.494	1	0.063	0.643	0.360	0	32.07***
583	1118	60	0.243	0.213	1.471	0	0.261	0.821	0.085	0	32.02***
1483	1023	43	0.062	0.207	0.628	1	0.279	0.786	0.060	0	31.49***
198	1038	67	0.409	0.177	2.078	1	0.090	0.607	0.060	0	31.24***
2303	1118	11	0.196	0.205	1.108	3	0.189	0.500	0.055	0	30.93***
2639	1054	152	0.080	0.110	0.870	2	0.378	0.357	0.050	1	29.39***
1079	1023	56	0.081	0.181	0.705	2	0.288	0.750	0.060	0	28.7***
3155	1019	21	0.108	0.237	0.587	4	0.261	0.536	0.060	0	27.86***

As can be seen in Table 3.8, many of the outlier interviews have a suspicous interview duration value. The interviews with respondents 2303 and 3155 took only 11 and 21 minutes, respectively. It seems impossible to finish the interview in that time. The interviews with respondents 680, 198, on the other hand, have a suspiciously small value for the IV *Filter*. What is more, the interviews with respodents 680, 2639 and 1079 have considerably low values for the IV *ERS*. Altogether, there seem to be some very suspicous interviews.

Furthermore, as can be seen in Table 3.7, we have considerably more outliers on the 1% level as would be expected among 1765 respondents. Apart from that, the \(\chi_6^2\)-distribution provides again a good approximation to the squared MaDi values (Figure 3.2).

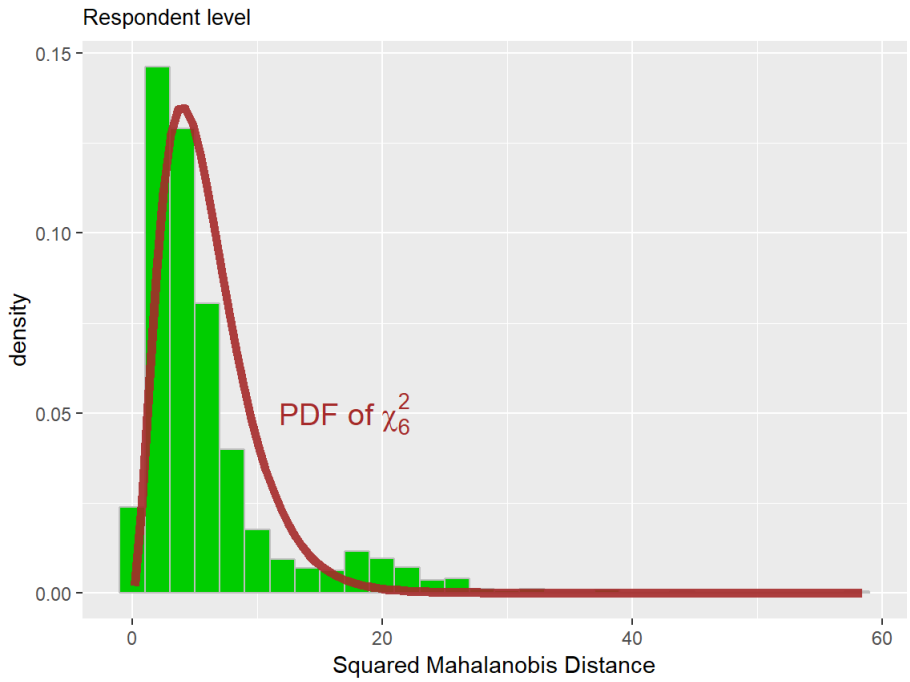

Figure 3.2: Histogram, Mahalanobis distance respondent level

Table 3.8 also lists the ID of the interviewers who conducted the interviews that were assigned high MaDi values. Yet, only two of these interviewers, namely 1038 and 1050, appear in Table 3.4, too. Thus, interviewers who conducted interviews with deviating IV values do not necessarily stand out on interviewer level. This is not surprising, considering that the IV values on interviewer level are obtained by averaging over all interviews that were conducted by the same interviewer. Only one or two deviating interviews can thus be balanced out by other interviews, if enough of them were conducted.

To see how the interviews conducted by ARIs fared on respondent level, Table 3.9 contains the MaDi score ranks of the two most deviating interviewers: 1004, 1038 (we did not consider interviewer 1058, because of his high *Filter* values)

Table 3.9: Respondent levels ranks of most extreme outlier interviewers

	Resp_ID	Int_ID	Minutes	ERS	MRS	VAR	Rounding	Filter	OQ_INR	Other_INR	TELE	Mah_Dis
Interviewer 1004												
58	2231	1004	58	0.069	0.577	1.298	1	0.171	0.500	0.060	0	21.59***
59	2228	1004	64	0.087	0.627	1.065	0	0.135	0.643	0.065	0	21.53***
129	2249	1004	73	0.111	0.430	1.311	0	0.432	0.321	0.050	1	17.65***
Interviewer 1038												
11	198	1038	67	0.409	0.177	2.078	1	0.090	0.607	0.060	0	31.24***
36	214	1038	108	0.224	0.109	1.509	1	0.297	0.393	0.060	0	23.86***
67	199	1038	65	0.130	0.178	0.805	4	0.288	0.393	0.050	0	21.11***
103	653	1038	50	0.267	0.320	1.126	3	0.216	0.500	0.055	0	18.95***
491	219	1038	59	0.327	0.179	1.419	3	0.171	0.643	0.060	1	6.31
923	659	1038	69	0.089	0.205	0.906	2	0.144	0.607	0.065	1	4.04
1551	210	1038	48	0.134	0.215	0.779	2	0.252	0.500	0.060	1	1.64

We can see that even though the interviews submitted by interviewer 1004 are not among the 15 interviews with the highest MaDi values, all of them are on high ranks, resulting in his high MaDi score. Interviewer 1038 on the other hand, has among his seven submitted interviews four that are on notably high ranks, the three others being rather unsuspicious.

3.2.3 Our Suggested approach

As mentioned earlier, having some very unusual interviews among 1765 is the norm and does not necessarily suggest that these interviews were fabricated. Our suggestion would thus be to combine the information of interviewer and respondent level and conduct a back-check on those interviews that are on a high rank (the threshold rank is up to the conductor of the analysis) and were conducted by ARIs. In our case, we would thus suggest to start reinterviewing with respondents 58, 59, 129, 11, 36, 67, 103 and to continue with the highly ranked interviews of the next ARIs. Since we do, however, not know which interviews were actually fabricated, we cannot show if that approach really gives good results.

We therefore believe that more research is necessary on which approach should be taken when the number of interviews varies greatly among interviewers and some interviewers have conducted so few interviews that their honesty is hard to assess. Future studies, which have a sufficient number of interviews and are applied to data where the genuineness status of each interview is known, could investigate if combining the information on interviewer and respondent level can deliver better results than considering only one level.

4 Limitations and Conclusion

4.1 Methodical Limitations

The MaDi is a powerful method to detect multivariate outliers. However, it also has its drawbacks. There are cases in which the MaDi would fail to properly detect outliers. Figure 4.1 illustrates such a case.

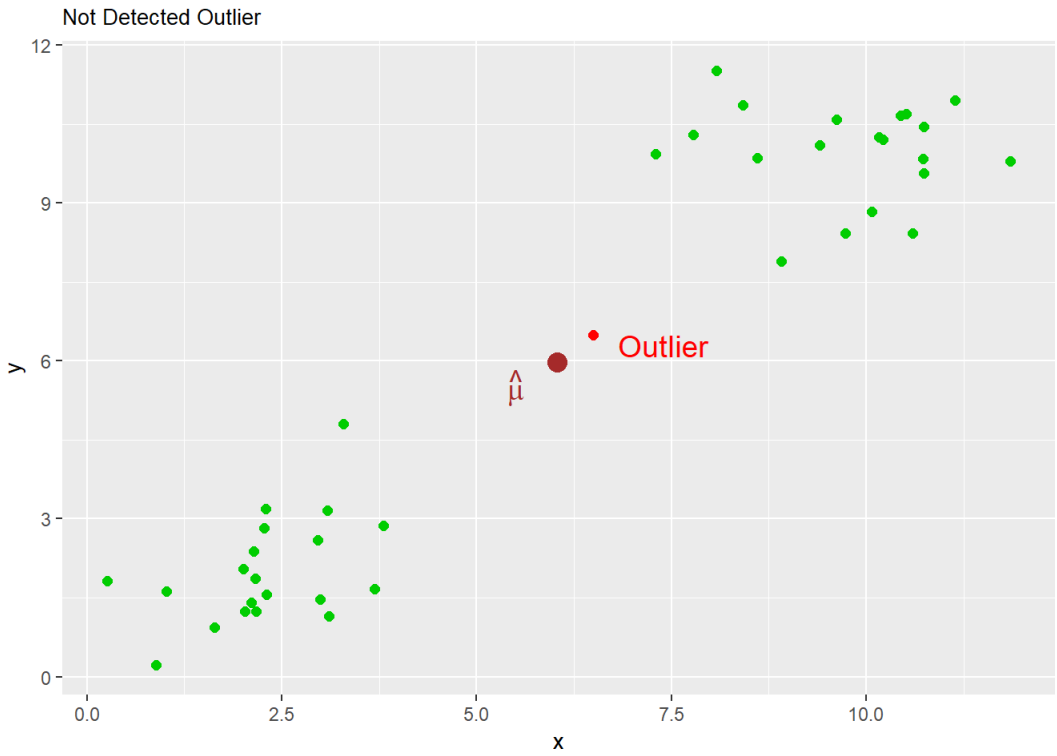

Figure 4.1: Example of a case in which the Mahalanobis distance does not detect an outier

Even though the red data point is clearly an outlier, it would have by far the smallest MaDi of all observations, since it is closest to the data center \(\mu\). Such cases can occur when several variables follow a mixed distribution, so that they have a “hole” when plotted. Our data, however, do not suggest any of those characteristics.

What is more, it needs to be stressed that our method relies on the assumption that most interviews are not fabricated. If, in fact, most of the interviews were faked, we would most likely identify the honest interviews/interviewer as outliers. Since the sample size is with 1765 interviews conducted by 138 interviewers reasonably large, we believe, however, that this is in all likelihood not the case.

Lastly, the ESS contains a lot of other information, especially in the contact form, that has the potential to be used as powerful IVs. We did not use that information, since our goal was not invent new IVs, but to apply already known IVs together with novel statistical methods. Future studies could, however, make use of this information.

4.2 Conclusion

In this paper, we have introduced using the Mahalanobis distance to identify possibly falsified interviews and/or fraudulent interviewers in the ESS 2016 Belgium. We first explained which indicator variables we selected and how their values were obtained, using only those that were already used in previous studies. We processed the data, obtaining an interviewer and respondent level dataset, on both of which we applied two principal component analyses, so that the obtained results were not biased in the direction of one indicator variable concept. We then elucidated the theory behind the Mahalanobis distance and applied it to our processed data. We were in this way able to filter out interviews that would be our preferred target for reinterviewing. Lastly, we illustrated limitations of our approach and argued that more studies in a supervised manner are necessary to ultimately show if our approach can yield satisfactory results.

5 References

Biemer, Paul P.; Groves, Robert M.; Lyberg, Lars E.; Mathiowetz, Nancy A.; Sudman, Seymour (Eds.) (2004): Measurement Errors in Surveys. Hoboken, NJ, USA: John Wiley & Sons, Inc (Wiley Series in Probability and Statistics).

Blasius, Jörg; Friedrichs, Jürgen (2012): Faked Interviews. In Samuel Salzborn, Eldad Davidov, Jost Reinecke (Eds.): Methods, Theories, and Empirical Applications in the Social Sciences. Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 49-56.

Bredl, S., Kötschau, K., & Winker, P. (2012): A statistical approach to detect interviewer falsification of survey data. In Survey Methodology (38), pp. 1-10. Available online at https://pdfs.semanticscholar.org/2acf/73730da1db6660945c11a12c373cc32b9c48.pdf, checked on 9/9/2018.

Bushery, J., Reichert, J., Albright, K. and Rossiter, J. (1999): Using Using date and time stamps to detect interviewer falsification. In American Statistical Association, 316-320. Available online at https://ww2.amstat.org/sections/srms/Proceedings/papers/1999_053.pdf, checked on 9/9/2018.

Christin Schäfer, Jörg-Peter Schräpler, Klaus-Robert Müller (2004): Automatic Identification of Faked and Fraudulent Interviews in Surveys by Two Different Methods. DIW Berlin. Available online at https://ww2.amstat.org/sections/srms/Proceedings/y2004/files/Jsm2004-000257.pdf, checked on 9/9/2018.

European Social Survey (2018): Frequently Asked Questions. Methodology - How long is the ESS questionnaire. European Social Survey.

Available online at https://www.europeansocialsurvey.org/about/faq.html, checked on 9/9/2018.

European Social Survey 8 - 2016 (2016a): Data Protocol. European Social Survey. Available online at https://www.europeansocialsurvey.org/docs/round8/survey/ESS8_data_protocol_e01_4.pdf.

European Social Survey 8 - 2016 (2016b): ESS8 - 2016 Data Download. European Social Survey. Available online at https://www.europeansocialsurvey.org/data/download.html?r=8, checked on 9/26/2018.

European Social Survey 8 - 2016 (2016c): Source Questionnaire. European Social Survey. Available online at https://www.europeansocialsurvey.org/docs/round8/fieldwork/source/ESS8_source_questionnaires.pdf.

European Social Survey 8 - 2016 (2016d): Variables and Questions. European Social Survey. Available online at https://www.europeansocialsurvey.org/docs/round8/survey/ESS8_appendix_a7_e02_0.pdf, checked on 9/26/2018.

European Social Survey 8 - 2016 (2018): Documentation Report. THE ESS DATA ARCHIVE Edition 2.0 The ESS Data Archive Edition 2.0. European Social Survey. Available online at https://www.europeansocialsurvey.org/docs/round8/survey/ESS8_data_documentation_report_e02_0.pdf, checked on 9/9/2018.

Forsman, Gösta; Schreiner, Irwin (2004): The Design and Analysis of Reinterview: An Overview. In Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, Seymour Sudman (Eds.): Measurement Errors in Surveys. Hoboken, NJ, USA: John Wiley & Sons, Inc (Wiley Series in Probability and Statistics), pp. 279-301.

Hood, C. and M. Bushery (1997): Getting more bang from the Getting more bang from the reinterviewer buck: Identifying ‘At risk’ interviewers. In American Statistical Association, pp. 820-824. Available online at http://ww2.amstat.org/sections/SRMS/Proceedings/papers/1997_141.pdf, checked on 9/9/2018.

Mahalanobis, P. C. (1936): On the generalised distance in statistics. In Proceedings of the National Institute of Sciences of India, pp. 49-55. Available online at https://insa.nic.in/writereaddata/UpLoadedFiles/PINSA/Vol02_1936_1_Art05.pdf, checked on 9/9/2018.

Menold, N.; Kemper, C. J. (2014): How Do Real and Falsified Data Differ? Psychology of Survey Response as a Source of Falsification Indicators in Face-to-Face Surveys. In International Journal of Public Opinion Research 26 (1), pp. 41-65. DOI: 10.1093/ijpor/edt017.

Murphy, J., Baxter, R., Eyerman, J., Cunningham, D. and Kennet, J. (2004): A system for detecting interviewer falsification. Phoenix, Arizona.

Schäfer, C., Schräpler, J., Müller, K. and Wagner, G. (2005): Automatic identification of faked and fraudulent interviews in the German SOEP. In Schmollers Jahrbuch (125), 183-193. DOI: 10.1007/s101820500188.

Swanson, D., Cho, M. and Eltinge, J. (2003): Detecting possibly Detecting possibly fraudulent data or error-prone survey data using Benford’s law. In American Statistical Association, 4172-4177. Available online at https://www.bls.gov/ore/pdf/st030020.pdf, checked on 9/9/2018.

Thissen, M. Rita; Myers, Susan K. (2016): Systems and processes for detecting interviewer falsification and assuring data collection quality. In SJI 32 (3), pp. 339-347. DOI: 10.3233/SJI-150947.

Warren, Rik; Smith, Robert F.; Cybenko, Anne K. (2011): Use of Mahalanobis Distance for Detecting Outliers and Outlier Clusters in Markedly Non-Normal Data: A Vehicular Traffic Example. Dayton, Ohio. Available online at http://www.dtic.mil/dtic/tr/fulltext/u2/a545834.pdf, checked on 9/9/2018.