Interim Dataset Analysis, Round 11

Report of Standardised Analysis for Detection of Interviewer-Controlled Issues

ESS ERIC Core Scientific Team[1]

March 20, 2023

[1]Authors: Roberto Briceno-Rosas (GESIS Leibniz Institute for the Social Science in Mannheim, Germany), May Dousak (University of Ljubljana, Slovenia), Paulette Flore and Joost Kappelhof (SCP - The Netherlands Institute for Social Science Research, The Netherlands) / For internal distribution only.

Contents

Reader's Guide **3**
0.1 Theorical Background . . . 3
0.2 Methods . . . 4
0.3 Evaluation of Results . . . 5
0.4 Further Investigation . . . 5
0.5 Implications . . . 6

1 Interviewers in the Netherlands **7**

2 Interview timestamps **9**
2.1 Interview duration and interview speed . . . 9
2.2 Interview time . . . 15

3 Item-Nonresponse **19**
3.1 Item-nonresponse for the variable income . . . 19
3.2 Item-nonresponse for Module H . . . 23

4 Observed variance between answers of respondents **26**
4.1 Variance between different interviews: Near duplicates . . . 26
4.2 Low observed variance within interviews: Non-differentiation . . . 27
4.3 Variance within and between interviews: Difference in latent variables . . . 30

References **35**

List of Figures

1.1 Distribution of interviews per interviewer 8
2.1 Distribution of interview duration by interviewer 10
2.2 Average interviewing speed for Module H 13
2.3 Time interval between consecutive interviews by interviewer 18
3.1 Proportion of item-nonresponse in income 20
3.2 Proportion of item-nonresponse in income with labelled outliers 22
3.3 Proportion of average item-nonresponse on items in Module H 23
3.4 Mean proportion of item-nonresponse on items in Module H with labelled outliers . 25
4.1 Percent Match Rates within interviewers 27
4.2 Mulligan scores for Module H per interviewer 29
4.3 Mean of component scores across interviews 31
4.4 Standard deviation of component scores across interviews 33

Reader's Guide

This report presents the results of the analysis of the interim dataset for the Netherlands. It aims to help national teams detecting undesirable interviewer behaviour in a timely manner. The report has been standardised for implementation in all participating countries as remote analysis of the interim dataset without the intervention of the international ESS team and without having to move securely stored data from the national level. The report focuses on indicators expected to signal issues with undesirable interviewer behaviour but it can also help detect other issues regarding data collection which are not related to interviewer behaviour. The reader's guide explains how to best utilize the report and take the appropriate steps for data colleciton. It is highly encourage to keep close communication with the CST via the Country Contact for discussion of results and consultation regarding any important decision. Please share the report with your Country Contact. This report should also be stored and deposited for documentation purposes along the data and other respective documents to the ESS Archive after fieldwork has been completed.

0.1 Theorical Background

Interviewers can have a systematic and substantial impact on the resulting data. Sometimes, their impact is beyond their control, for example, respondents might answer questions somewhat differently when interviewed by a male interviewer compared to when interviewed by a female interviewers. Other times, their impact is well within their control and it is a direct consequence of their behaviour. When analyzing the interim datasets, we want to focus on interviewer-controlled issues, which can be corrected for during the fieldwork if necessary.

Types of interviewer-controlled issues

Interviewer-controlled issues varied depending on the level of control an interviewer has over it and the causes of the issue. One classification used in the ESS to gain a better understanding of the different ways undersirable interviwer behaviour can occurred is provided by Stoop et al (2018):

- Undesirable interviewer behaviour driven by context: Some examples are speeding through the interview when the respondent seems close to breaking off the interview, chatting with respondents who are unsure about their answers, paraphrasing a question when the respondent misunderstands, etc. In these cases, the interviewer depart from the expected behaviour due to an external element, for which the interviewer has little control of. They can be considered unavoidable to some extend, although in some situations the reaction of the interviewer could be improved to meet the desirable behaviour. More importantly, these issues should not be systematically related to the interviewer, but case-specific instead.
- Undesirable interviewer behaviour that could be avoided: Some examples of these type of behavior are speeding through the interview, skipping introductions to questions, forgetting to hand over showcards, not properly reading answer categories, etc.
- Unintentional errors: like erroneously interviewing the wrong person, recording the wrong day for the interview, keying the wrong answer category, etc.
- Deliberate falsification: curbstoning (creating the answers for an entire or partial interviews), partially duplicating interviews (copying answers), selecting available household members as respondents instead of a random member of the household because they are more cooperative or more often at home when the interviewer contacts the sample unit, incorrectly recording

answers to filter questions to reduce the duration of the survey, etc. In the cases, the behaviour is intentional and constitute a depature from the standards of survey interviewing set by the ESS.

Understanding the type of interviewer-controlled issues is important for deciding the course of action. For example, an interviewer who mistakenly missed parts of the questionnaire might require better training or briefing on the ESS questionnaire, while an interviewer who skipped parts of the questionnaire to save time might need to be excluded from the fieldwork activities. Data analysis can provide clues about the type of issues and help focus the monitoring efforts. However, further investigation is most likely necessary to assert the type of behaviour that caused the issues.

Prevention and detection

One of the key steps for successful promotion of desirable interviewer behaviour is detecting issues in the field as early as possible. Detection leads to prevention of further issues and therefore to mitigation of the impact of undesirable interviewer behaviour. Even if issues can not be corrected, understanding the causes that lead to issue allows to make better informed decision and document problems for data users. The more time it has passed between the detection of a data issues and the fieldwork, the harder it is to understand the issue.

Strategies for dectecting interviewer-controlled issues

Data analysis of interim datasets is one out of multiple strategies that can be used to detect interviewer-controlled issues during data collection. Other strategies include observation of interviewer behaviour in the field, interviewer debriefing, or back-checks or recontact sample units, analysis of fieldwork paradata (e.g. contact forms data), or activity tracking mechanism. Therefore, the interim dataset from the main questionnaire is provides “one piece of the puzzle” when attempting to reconstruct fieldwork activities in the process of data collection and understand.

Standardised v country-specific analysis

Standardised data analysis provides a basis for detection of issues related to interviewer issues. It also allows international surveys like the ESS to establish a minimum quality benchmark that is applied in all participating countries and serves to improve comparability across countries. However, standardised analysis for all participating countries cannot account for country-specific characteristics that are particular to national or regional context, specific survey design characteristics, cultural background, technical particularities of the tool used for the national teams for data collection o country-specific fieldwork decision. It is highly encourage to interpret, adapt and conduct further analysis that would address particular concerns expected in each country.

0.2 Methods

Areas of Analysis

Three areas of analysis have been defined in this report. The first section presents the results on the analysis of the timestamps from the interview. The second section focuses on item non-response. The third section investigates response patterns of respondents and how these related to allocations within interviewers.

Quality Benchmarks (flags)

The quality benchmarks, also known as flags, function as default threshold for each indicators presented in the report. Their purpose is focusing the attention to possible issues rated to interviewer behaviour. These benchmarks are arbitrary and they has been defined based on experience of fieldwork activities across participant countries. The thresholds are by no means fixed and they can be revised to meet country-specific needs and fieldwork characteristics.

For the interim dataset, quality benchmarks have been defined with a low sensitivity, meaning that is expected that higher number of cases will be flag without an actual problem having occurred (false positives). The flag aim is to raise attention about possible issues for further investigation, instead of reporting the existence of an issue *per se*.

0.3 Evaluation of Results

Figures, Tables and Detailed Tables in Annex Folder

In the report, results are presented in figures and table to provide a quick overview of the issues. Promatic cases or interviewer are usually highlighted. However, it is not always possible to provide a adaquate overview of all issues for all countries. Therefore, indicators are saved into the Annex folder together with the report for further invatigation.

Adjustment of Quality Benchmarks

Adjustment of the quality benchmarks might be necessary to better tune to sensibility to country-specific context or to conduct a more focused investigation of issues. There are two ways to adjust the quality benchmarks used as default for this report: (1) recoding of the

0.4 Further Investigation

All flags provided by this report are granted further consideration and investigation with the goal of gaining understanding about the causes of the issues. The steps and efforts will depend on the type of flag and the related level of concern.

Contextualization

National teams and survey agencies have access to important information about fieldwork activities, which this report does not account for. For example, prior issues with specific interviewers or details about specific geographical area of sample units can substantially influence the levels of concern raised by flags. Therefore, the first step for the national teams should be to contextualize the reported flags with information available about fieldwork, including the characteristics of the interviewers, the sample units, the details of the CAPI-system, recent fieldwork activities, etc. The information required for contextualization will depend on the respective quality benchmarks and possible explanation for their occurrence. For example, the expected travel distance between two sample units can provide crucial information when asserting the feasibility of very short interval of time between two interviews from the same interviewer.

Debriefing of Interviewers

An important step in clarifying the issues is to query interviewer about possible explanation of the issues observed. Interviewers are one of the most important source of information for explaining issues observed in the data as they were present in the production of the data. They can help reconstruct the

fieldwork activities that lead to the issues observed. They should also be seen as partners in solving the issue as their future behaviour might help avoid problem. We recommend that this interaction takes place in a constructive and respectful manner as it. We recommend to document the debriefing results in a short minute that could help later in the assessment, especially if back-checks are conducted.

Back-checks

Please follow the ESS guidance on conducting the back-checks. Please note that when conducting back-checks with the aim of clarifying specific interviewer-related issues, you might need to ask the sample units about specific information that clarifies the issues. For example, if there are concern about parts of the questionnaire not having been asked, you might need to ask the respondent whether they recall being asked about those specific topics and present one or more of the questions to help them recall.

0.5 Implications

Course of Action, Discussion and Documentation

Once an understanding of the issues has been establish, it is time to decide and discuss the course of action. The course of action should always be decided on the case-by-case approach, considering all available information and with the discussion with the respective stakeholders. Some example of course of action might involve management of the interviewers, like for example re-briefing, re-training, supervised interviewing exercises, or removal of interviewers from field activities. Management of the interviewers are under the responsibility of the national team and survey agency, however, consultation with the CST is encouraged. Other solutions might require a adjustment of the survey design or manipulation of the data, like recontacting sample units, correction of erroneous data, removal of data (partial or complete), redrawing the sample. Please note that any course of action that involved the modification of the fieldwork plans, contact strategy, manipulation of data, or sampling design are not to be taken lightly and need to be consulted with the Core Scientific Team before any action is taken.

In some cases, the course of action might just be the documentation of the issues, their probable cause, and the step taken to clarify the issues. This allows the ESS to provide explanation to data users about possible issues they might spot when making use of the data. It is recommended to document any course of action taken as a results of quality control and flags raised by this report.

1 Interviewers in the Netherlands

The data provided to the tool contained a total of 53 interviews, of which 53 contained a valid interviewer number. It is assumed that all interviews have been conducted by an interviewer. Only cases with a valid interviewer number are analyzed in this report.

A total of 10 interviewers have at least one completed interview. Before interpreting the quality indicators in this report, it is important to bear in mind the number of interviews each interviewer has completed. Table 1.1 provides a descriptive summary of the interviews per interviewer while figure 1.1 visualizes its distribution. The distribution of cases provide a rough picture of the fieldwork activities and an indication of the robustness of the results in this report.

Table 1.1 Descriptive summary of interviews per interviwer

min	max	mean	sd	Q1	Q2	Q3	n
2	10	5.3	2.3594	4	5	6	10

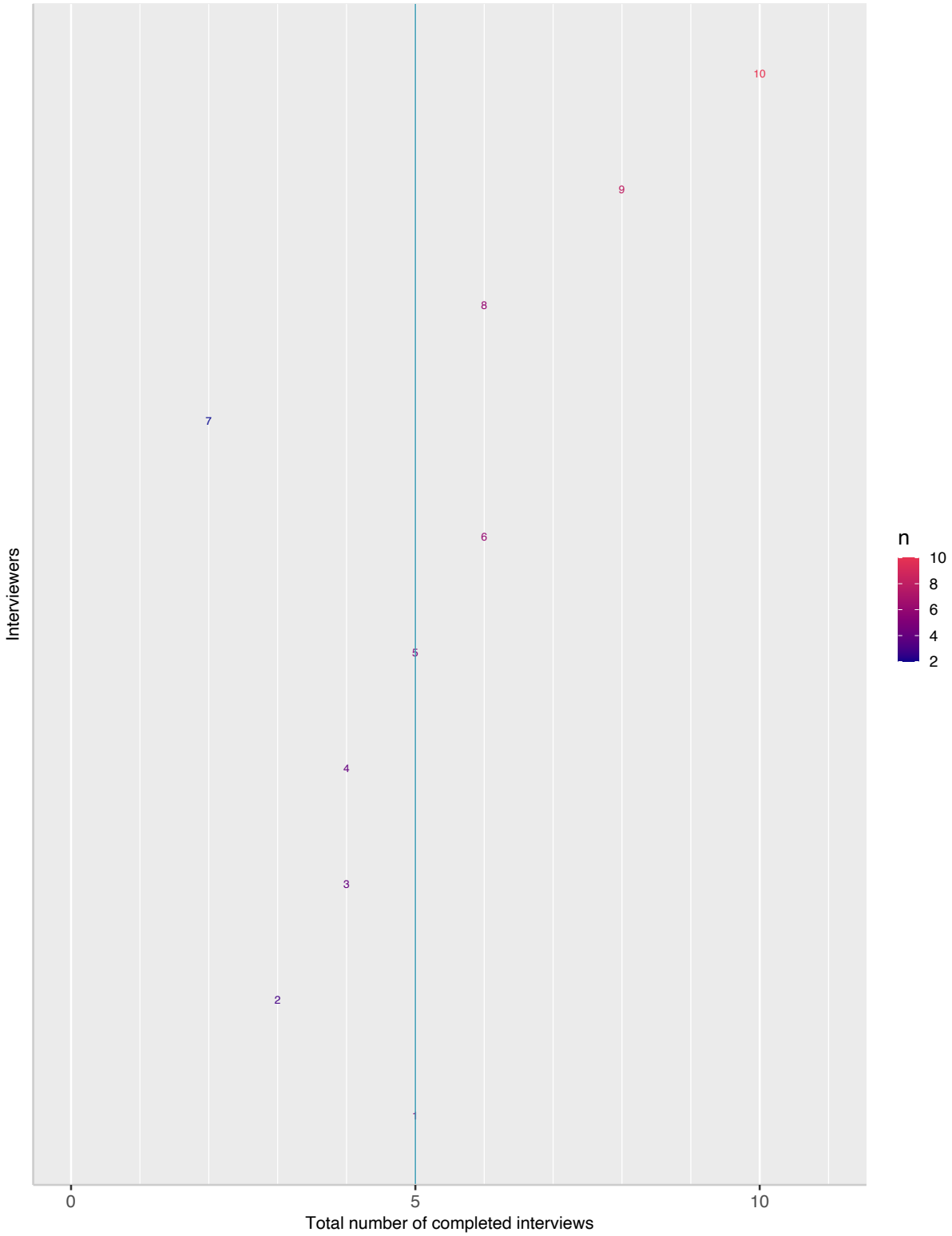


Figure 1.1 Distribution of interview per interviewer

Note: The vertical blue line represents the median number of completed interviews

2 Interview timestamps

In this section, indicators for detection of issues in interviewing process are drawn from the analysis of timestamps recorded for each interview (start and end of the interview, including timestamps for each module, and date and time of interview). Timestamps provide critical information about the conduction of the interview in the field and can help detect issues in the interviewing process.

First, we look at the interview duration and interview speed on both the interview level and module level. Second, we analyze the time of the interview relative to the time of other interviews from the same interviewer. Please note that any issues highlighted require further investigation by the national teams.

Countries that follow the recommendation of including more detailed timestamps (e.g. timestamps per item) into their CAPI systems are able to conduct further analysis on those indicators. The ESS DIB team can provide further assistance if needed.

2.1 Interview duration and interview speed

Interview duration refers to the length of the interview excluding optional country-specific questions, the interviewer questions, and general administration of the contact procedures. Interview speed refers to the average speed (in terms of question per minute) in which the interview progresses. The average interview speed is calculated as the number of applicable questions for a specific respondent divided by the duration of the interview. It can be calculated at the interview level or another level, such as the module level (see Section 2.1.2).

Both duration and speed should be considered when assessing data quality. For example, extreme outliers in interview duration and/or interview speed can indicate a deviation from interviewing protocols (e.g. error in the recording of timestamps) or even falsification of interviews.

2.1.1 Unlikely interview duration

Very short or very long interviews can be attributed to interviewer behaviour, but also to technical problems (e.g., software problems), atypical interviews, or respondent behaviour (e.g., partial interviews) (see Vandenplas, Beullens, & Loosveldt, 2019, p. 252). Detecting outliers can help identify systematic issues related to interviewers as a whole. Therefore, it is important to consider all interviews conducted by interviewers that have produced interviews with an unlikely duration. As mentioned before, very short of very long interview durations (i.e., outliers) could indicate a potential problem and in this analysis we define short interviews as interviews having lasted less than 30 minutes and long interviews as interviews having lasted 180 minutes or more.

An overview of the distribution of interview duration by interviewer can be seen in Figure 2.1. Table 2.1 shows the frequencies of an interview with unlikely duration per interviewer. For more details, see table in the annex folder with the interview duration per case and interviewer (“Interview duration.csv”).

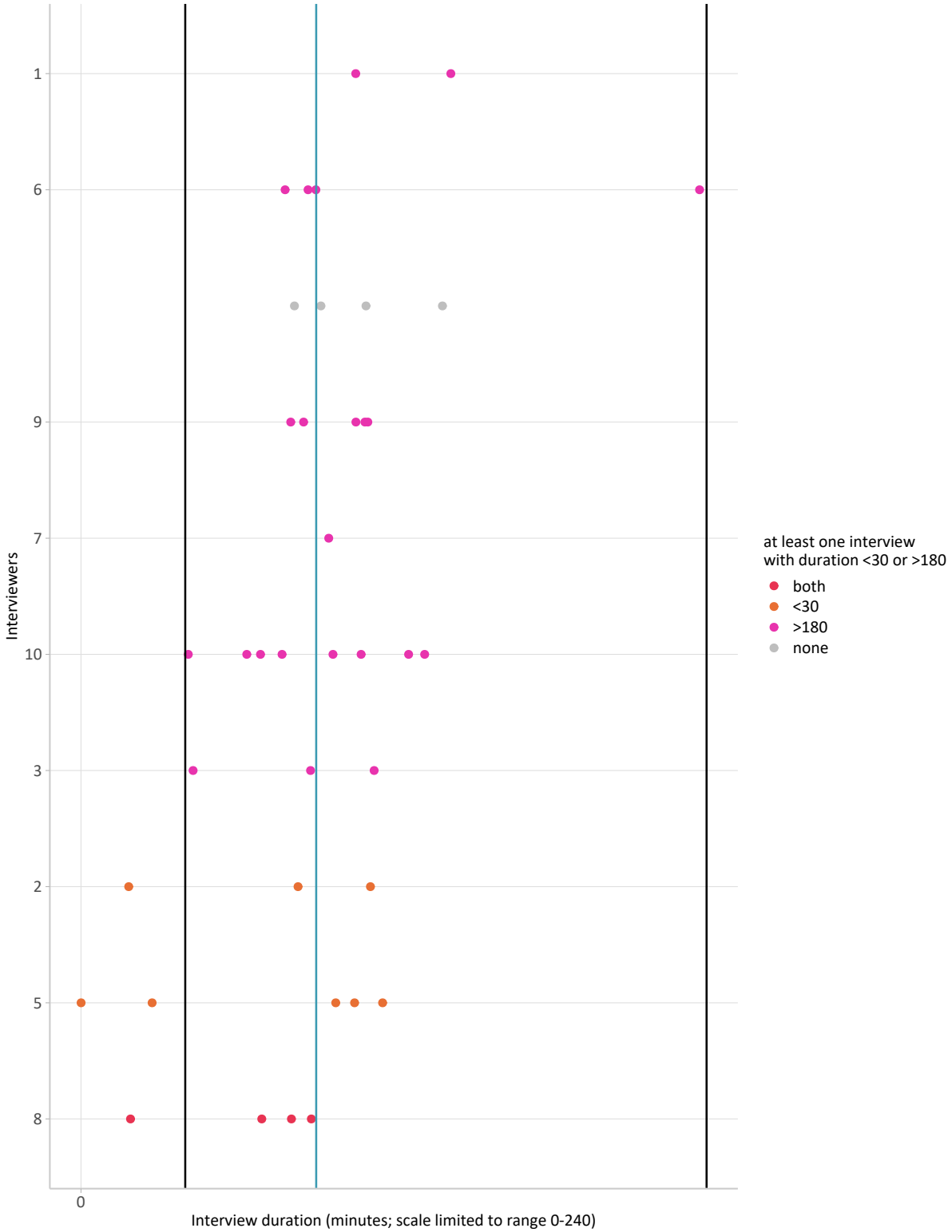


Figure 2.1 Distribution of interview duration by interviewer (limits 0-240 minutes)
Note: The figure shows only interviews with a duration of 0-240 minutes to give a better overview (see respective annex file for all interview duration records). Interviewers with at least one interview with a duration <30 or >180 minutes are labelled in the y axis with their interviewer ID. The vertical black lines represent the fences of unlikely interview duration (less than 30 and more than 180 minutes). The vertical blue line represents the mean of duration for all interviewers after deleting interviews with improbable durations.

Table 2.1 Interviewers with unlikely interview durations

Duration	Interviewer ID	Number of interviews
<30	2	1
	5	2
>180	1	3
	3	1
	6	2
	7	1
	9	3
	10	2
both	8	3

The general recommendation here is to go back to the survey agency (or fieldwork coordinator or interviewer in case of an in-house survey) and have them check these results with the interviewer in question. Especially in the case of interview duration, other indicators like speed and item non-response (see next sections) should be taken into consideration.

2.1.2 Interviewing speed

Previous research in the ESS has shown that the interview speed of interviews is linked to the impact interviewers have on the answers of the respondents, known as interviewer effects (Vandenplas, Beullens, & Loosveldt, 2019). Interviewer effects are measured by the extent to which the variance of items is explained by the allocation of cases within an interviewer. Vandenplas et al. (2019) showed that there are larger interviewer effects for slow and for fast interviews compared to moderate interview durations.

Due to filter questions in the ESS questionnaire, the number of applicable questions varies from interview to interview. Therefore, the speed of the interview provides a more comparable indicator across interviews. We calculate the speed over an interview v as the number of applicable items in that interview q divided by the time needed to complete the interview t in minutes for each respondent i:

$$v_i = q_i / t_i \tag{1}$$

Equation 1 indicates the number of items answered per minute. Interview speed can be calculated on the interview level and on the module level. Both can be informative when trying to assess the potential for undesirable interviewer behaviour. Below we provide you with an example as to how the estimated interview speed at the module level can aid us in detecting undesirable IB as well as help us in our understanding of potential mitigation strategies going forward.

Interviewing speed on the module level: Module H

Figure 2.2 shows the average interviewing speed for each interviewer and interviewers with at least four interviews for module H. We made this differentiation, because the probability that outliers in speed are incidental findings is lower for interviewers with at least four interviews. Interviews with unlikely interview durations (<30 minutes or >180 minutes) were excluded from the analysis because these are already detected in Section 2.1.1 as outliers. We define an interviewing speed 1.5 times out of the Interquartilerange (IQR) as outliers with very fast or very slow interviewing speed. Those interviewers are labelled in the figure and listed together with their average interview speed in Table 2.2. For more details, see table in the annex folder with the interview speed for module H per case and interviewer (“Interview speed module H.csv”).

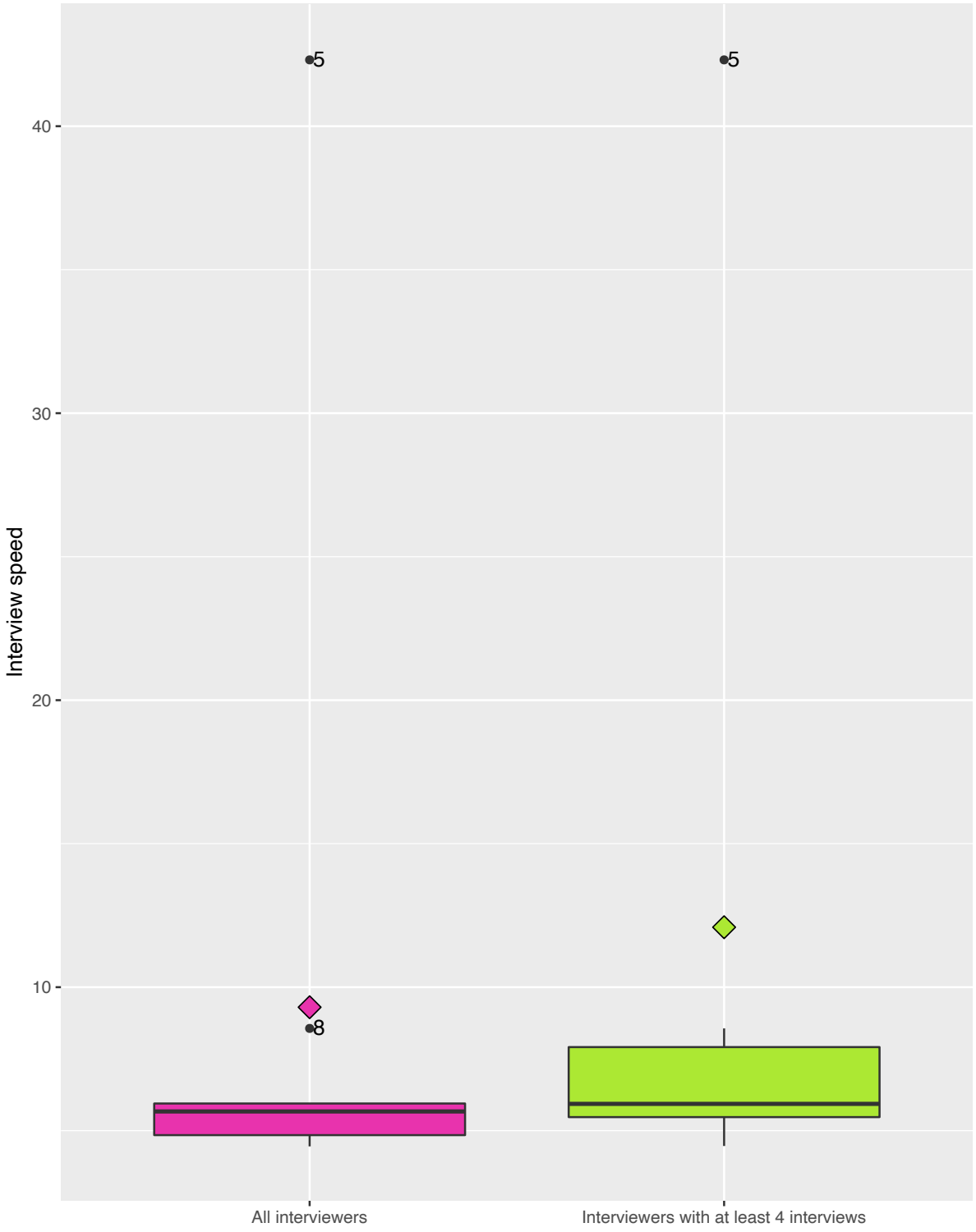


Figure 2.2 Average interviewing speed for Module H

Note: Interviewing speed is defined as items answered per minute. The squares represent the mean of the distribution. Interviewers that are > 1.5*IQR from the borders of the box are labelled.

Table 2.2 Interviewers with unlikely interview speed

Group	Interviewer ID	Average speed
All interviewers	8	8.563106
	5	42.311304
Interviewers with at least 4 interviews	5	42.311304

A very high interview speed should give rise to further investigation as to how these interviewers conducted these modules. It is advisable to go back and ask what exactly happened there. In general terms: Module H is the last questionnaire module and there could be some unfortunate, but valid reasons as to why some of the interviews were conducted with this average speed (e.g., respondent broke off the interview so the interviewer coded all remaining answers as 'no answer'). However, it is of course of particular interest if this pattern is observed multiple times within the same interviewer as this may indicate a more systematic undesirable behaviour such as rushing through the questions, not reading these questions to the respondent, or not challenging the respondent when he/she does not differentiate anymore.

2.2 Interview time

The (date and) time in which the interviews were conducted is another indicator that has proven useful for detecting undesirable interviewer behaviour or other issues in the field. It can inform about compliance with contacting and interviewing protocol, performance in achieving the respondent's cooperation, as well as error in the recording of timestamps. First, we look at interviews with an overlapping interview time conducted by the same interviewer. Second, the time of the day in which interviews were conducted is checked for unlikely interview hours (e.g., the middle of the night). Third, we look at the number of interviews conducted on the same day by the same interviewer. Lastly, the time between consecutive interviews of the same interviewer is analysed and its plausibility assessed.

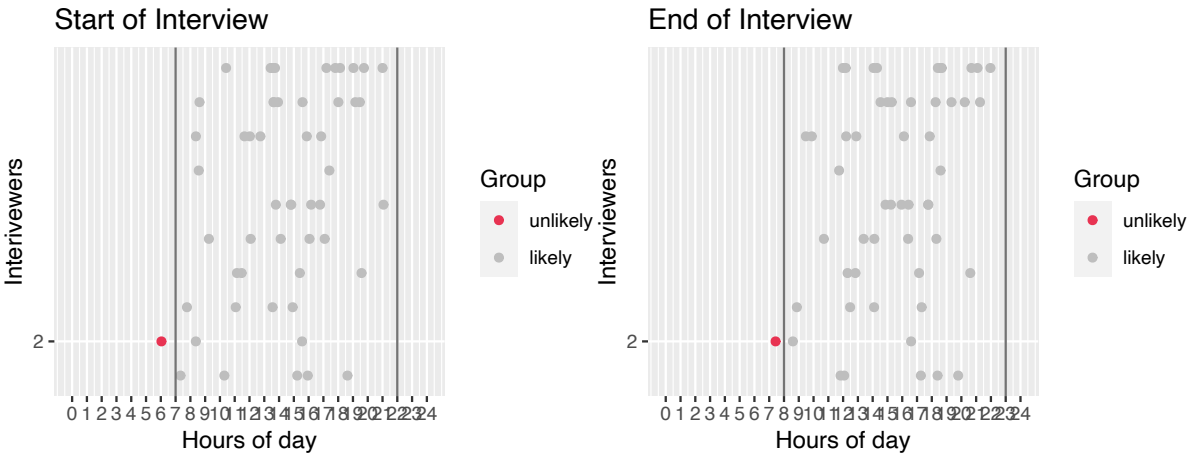
2.2.1 Overlaps of interview time

The overlap of the interview time of different interviews conducted by the same interviewer is an implausible outcome of fieldwork activities and could indicate the occurrence of poor interviewer behaviour. Overlapping times could be the result of interviewer deviating from interviewing protocol (e.g. entering the wrong interview time manually, forgetting to close the CAPI application, or multiple interviewers conducting interviews under the same ID). Moreover, it could be the result of interview falsification (e.g. fabrication of full interviews or parts of the interviews).

However, before jumping to these serious conclusions, it is important to realize that overlapping interview times can also be the result of technical CAPI issues. These are serious issues which should be corrected on time to avoid contamination or the loss of data, but they are not necessarily the result of undesirable interviewer behaviour. Still, if an overlapping of interview times is detected, it is necessary to look up the causes. The national team should investigate the cases in detail to clarify the reasons of occurrence. Please also inform the CST about the outcome of further investigation of these cases.

2.2.2 Time of interview

We expect most of the interviews to take place during daily hours of social life because that is when potential respondents would accept an invitation to conduct an interview as well as that these hours would correspond with the 'regular' working hours of an interviewer. We would view interviews with a starting time between 10 pm and 6 am and/or ending between 11 pm and 7 am as unlikely. It is possible to conduct interviews during these hours, but they should be looked into more detail. This indicator could inform about possible issues in the data collection or the recording of timestamps, but one should of course take notice of socially accepted interview times in a particular country. However, if multiple interviews from the same interviewer have taken place at an unlikely time, we highly recommend conducting back-checks on these cases.



The figure shows the distribution of time interviewer started and ended their interviews. For more details, see table in the annex folder with the time for the start and end of the interviews ("Interview start and end time.csv").

2.2.3 Number of interviews on the same day by a single interviewer

The maximum number of interviews in a single day by a single interviewer is an indicator of peak performance by an interviewer. Interviewers should organize and schedule their contact attempts and appointments in such a way that allows them to work as efficiently as possible. However, in some cases, this peak performance indicator can be related to non-compliance with contacting, selection or interviewing protocols. Therefore, national teams are recommended to closely monitor the work of interviewers with an extremely high number of completed interviews within a single day.

For more details, see table in the annex folder with the maximum number of interviews per day achieved by each interviewer ("Maximum interview per day.csv"). As mentioned before it is entirely possible to conduct this many interviewers on a single day. Please adjust this threshold of three interviews if is appropriate for your country. These adjustments should be discussed with the CST. If the threshold was reached for more than one case, we would recommend back-checking these interviewers if, in conjunction with the time interval indicator (see Section 2.2.4), the combination becomes improbable.

2.2.4 Time interval between consecutive interviews by a single interviewer

The time between consecutive interviews by a single interviewer could also be an indicator of non-compliance with contacting and selection protocol or poor interviewer behaviour. Similar to the number of interviews of on the same day, interviewer should organize and schedule their contact attempts and appointments in such a way that allows them to optimize the (travel)time between interviews. However, extremely short time intervals could also indicate issues affected by undesirable interviewer behaviour, especially if these occur multiple times within the workload of a single interviewer.

To this end, the analyses would show the minutes between the end of an interview and the start of the following interview within each interviewer. In many cases, the resulting number of minutes is very large, e.g., because the following interview took place another day. Whenever multiple interviews took place on the same day, it is more relevant to check the interval between two interviews.

Figure 2.3 lists interviewers with interviews for which the previous interview was conducted 120 minutes ago or less. Shorter and potentially critical intervals of less than 30 minutes are highlighted in orange to allow an easy check of patterns and clusters. As for most timestamp indicators, this figure can be a good starting point to check the underlying reasons with the survey agency and highlighted interviewers. For more details, see table in the annex folder with the time interval between interviews of the same interviewer (“Time interval between interviews.csv”).

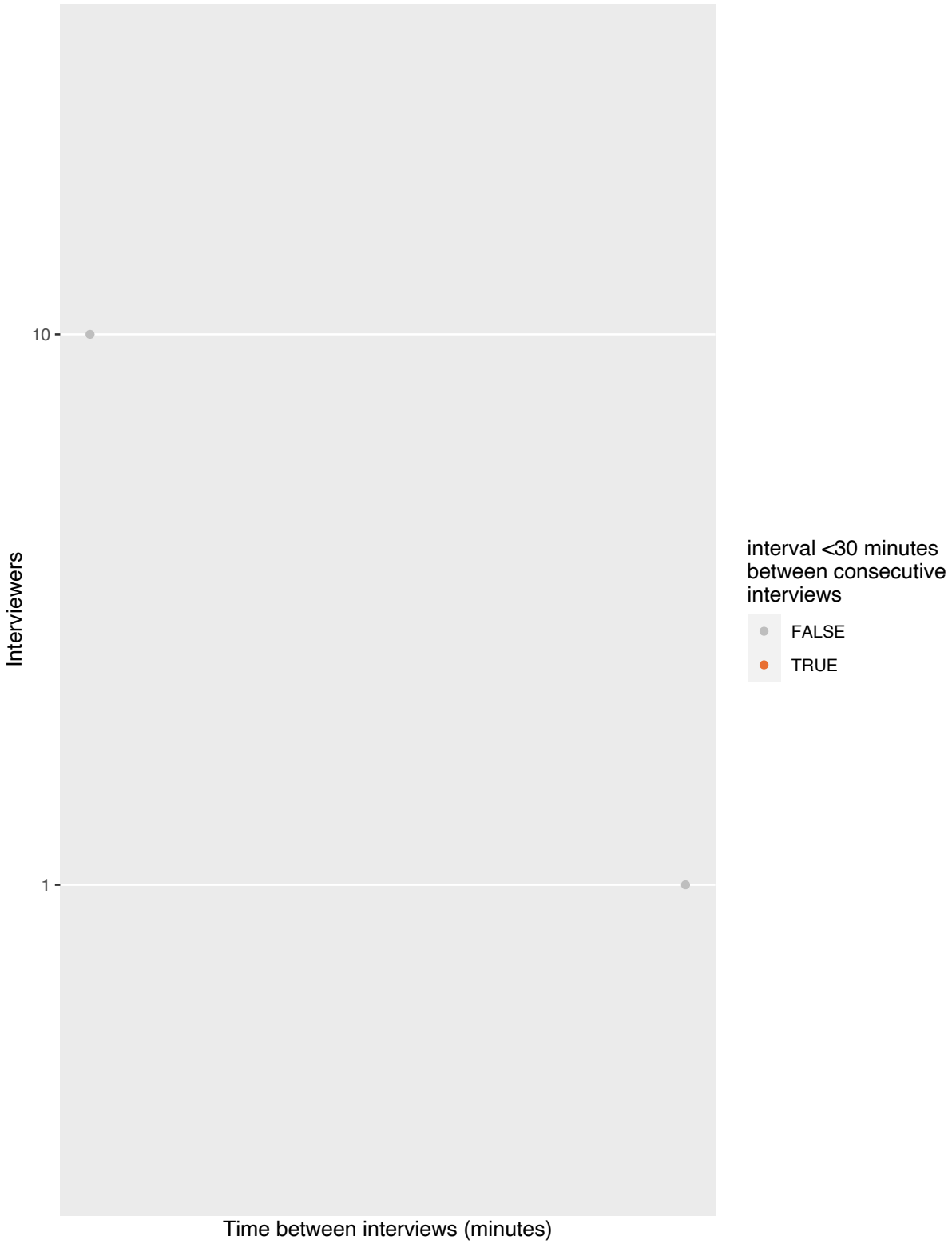


Figure 2.3 Time interval between consecutive interviews by interviewer
Note: For a better overview, only intervals <= 120 minutes were considered in the graph. Interviewer with at least one interval < 30 minutes are labelled.

3 Item-Nonresponse

This section focuses on item-nonresponse on the interviewer level for the participating ESS countries. In this section, we will take the variable income and module H to study (the proportion of) nonresponse per interviewer in the Netherlands. Nonresponse is in this section defined by either giving the item response "Refusal," "Don't know," or "No answer."

The proportion of nonresponse per interviewer is given by the number of the selected part of interviews that elicited a nonresponse divided by the total number of interviews held by the interviewer in question. Whereas a low or high proportion of item-nonresponse is not necessarily a sign of misconduct, sloppiness or any wrongdoing on the interviewer's part, it might be insightful to study item-nonresponse nonetheless. A reason for further investigation might be if the proportion of missing values within an interview is unusual in conjunction with other unusual patterns in the data for this particular interview, especially if this pattern is observed within multiple interviews conducted by the same interviewer. For instance, a high proportion of item-nonresponse combined with unusual interview duration might give us some reason to worry because this may indicate that an interviewer skipped sections of the questionnaire during the interview. We would like to warn NCs that in isolation an unusual proportion of item-nonresponse is not enough ground to flag an interviewer as a suspicious case.

In the remainder of this section, we will give an overview of average item-nonresponse, and item-nonresponse on the interviewer level for the selected variables.

3.1 Item-nonresponse for the variable income

The decision to zoom in on item-nonresponse for the variable *income* stems from the sensitive nature of this question. Discussing someone's income is taboo in many cultures and nonresponse on an item like this is typically high (Tourangeau & Yan, 2007). Therefore, it seems unlikely to find very low item-nonresponse on this item, and even more unlikely if this pattern is observed on many interviews conducted by the same interviewer. However, the degree to which this question elicits a nonresponse may vary from culture to culture. This should be considered when interpreting the results.

We start with an overview of the occurrence of item-nonresponse in the Netherlands, to obtain some insight on the degree to which item-nonresponse is common within the country. Then we proceed to investigate unusual patterns in the data regarding nonresponse on the *income* item between interviewers.

For an overview, we, firstly, check the percentages of nonresponse on the *income* item. Secondly, we summarize the proportion of item-nonresponse on the *income* item of all interviews conducted by the same interviewer. The proportion of item-nonresponse p on the *income* item for each interviewer j is calculated as follows:

$$p_j = inr_j / ni_j \tag{2}$$

where inr_j gives the number of interviews in which a respondent gave a nonresponse on the *income* item for interviewer j and ni_j the total number of interviews conducted by interviewer j.

Figure 3.1 shows the distribution of the proportion of item-nonresponse in the variable income, calculated according to Equation 2 separately for all interviewers and interviewers with at least 10 interviews.

Overall, we can expect the proportions of nonresponse per interviewer to be relatively low, but there might be some exceptions. Looking at all interviewers, we can expect some extreme proportions for interviewers who conducted only one or a few interviews. For a better overview, the distribution of interviewers with at least 10 interviews should therefore also be considered.

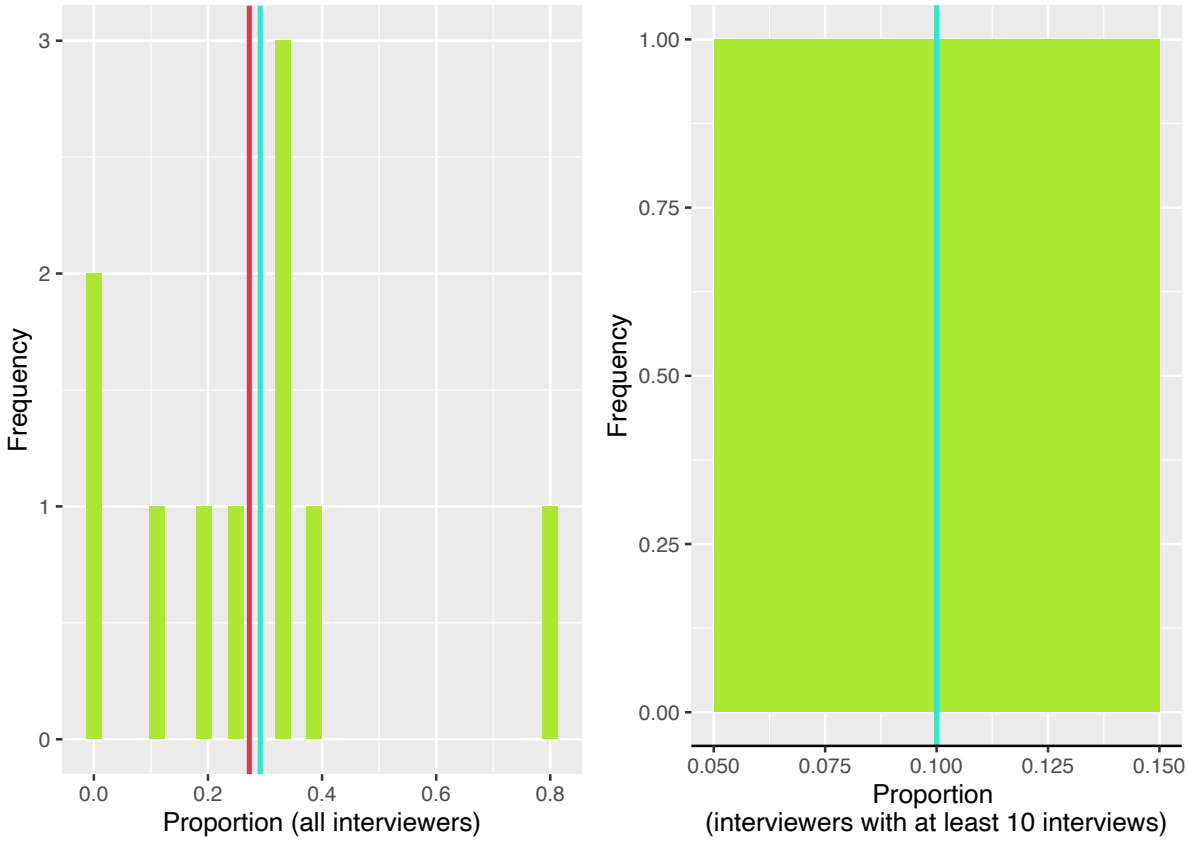


Figure 3.1 Proportion of item-nonresponse in income per interviewer
Note: The red line indicates the mean of the proportion. The blue line indicates the median.

For interviewers who conducted many interviews, we expect there to be some, but not a lot of nonresponses. To find unusual proportions of item-nonresponse, we need to zoom in on interviewers who conducted a fair number of interviews, and who have either a very low or a high proportion of nonresponse.

Figure 3.2 gives an overview of outliers in proportion of nonresponse, both for the complete sample and for the subset of interviewers who conducted 10 interviews or more. For more details, see table in the annex folder with the proportion of item nonresponse per interviewer for income ("Prop item nonresponse income.csv"). We particularly recommend looking at interviewers who have conducted 10 or more interviews and had a proportion of nonresponse larger than 50%. Because nonresponse is common for the variable income, very little or no nonresponse might also be unusual and worthy of our attention. Those interviewers might be investigated more, but should not be flagged solely on this data.

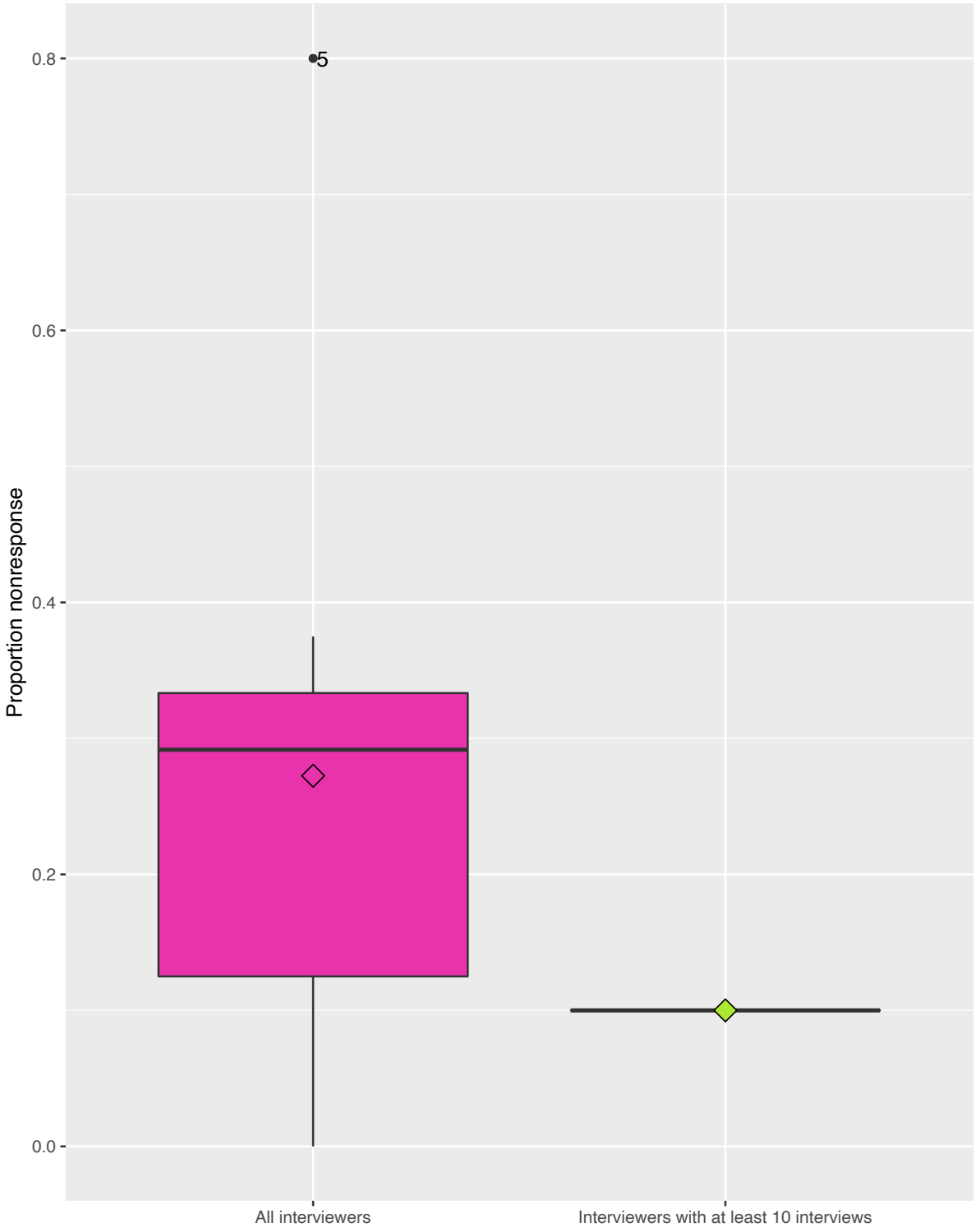


Figure 3.2 Proportion of item-nonresponse in income with labelled outliers
Note: The squares represent the mean of the distribution. Interviewers that are > 1.5*IQR from the borders of the box are defined as outliers.

3.2 Item-nonresponse for Module H

In addition to studying item-nonresponse for a sensitive question like income, we additionally aim our attention at the patterns of nonresponse in a module of less sensitive questions, in this case, *Module H* (21 items). Again, we first give an overview of item-nonresponse on the items in *Module H* in the Netherlands, and in the second part, we zoom in on unusual properties.

For an overview, we, firstly, check the percentages of nonresponse on the items in *Module H*. Secondly, we summarize the proportion of item-nonresponse on items in *Module H* of all interviews conducted by the same interviewer. The proportion of item-nonresponse p on items in *Module H* for each interviewer j is calculated according to Equation 2. Subsequently, for each interviewer, we calculate the mean proportion of item-nonresponse over these 21 items.

Figure 3.3 shows the distribution of the average proportion of item-nonresponse for all 21 items for each interviewer. It is calculated according to Equation 2 separately for all interviewers and interviewers with at least 10 interviews.

Overall, we can expect the proportions of nonresponse per interviewer to be relatively low, but there might be some exceptions. Looking at all interviewers, we can expect some extreme proportions for interviewers who conducted only one or a few interviews. For a better overview, the distribution of interviewers with at least 10 interviews should therefore also be considered.

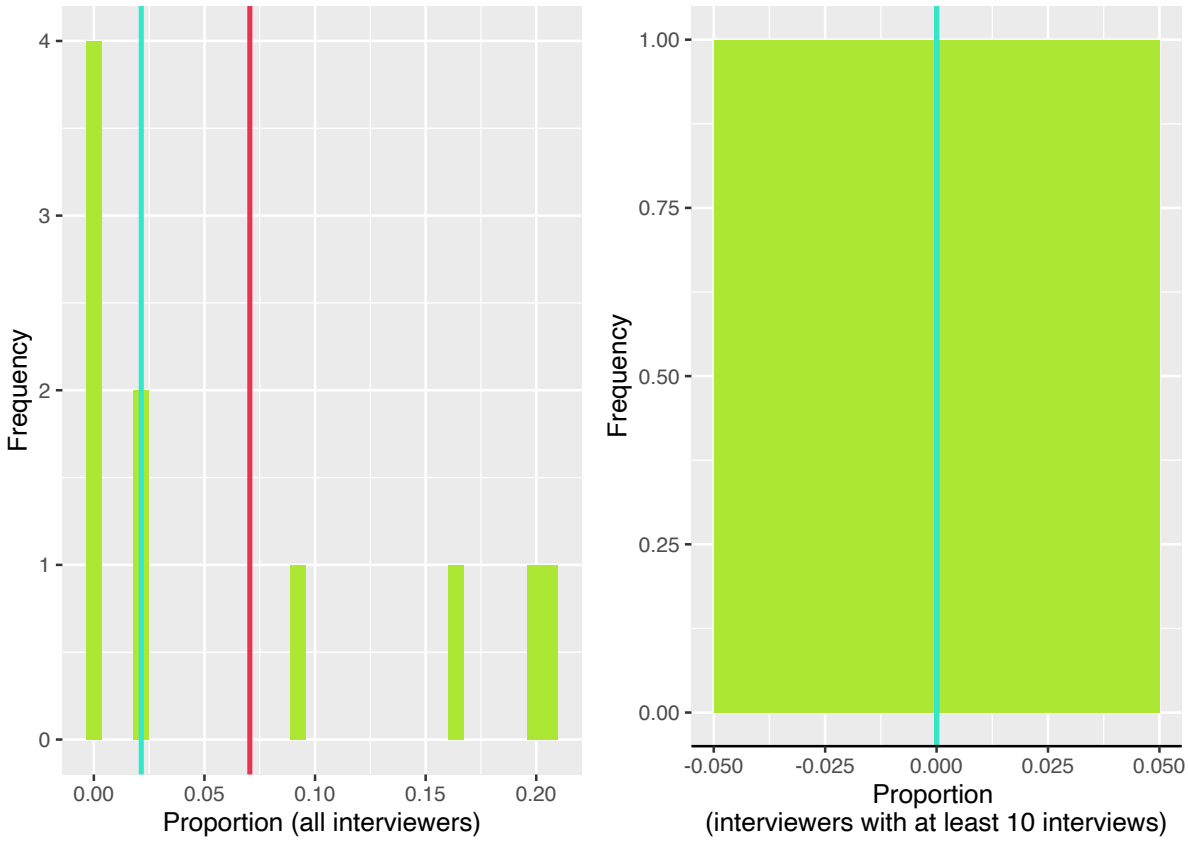


Figure 3.3 Mean proportion of item-nonresponse on items in Module H
Note: The red line indicates the mean of the proportion. The blue line indicates the median.

For interviewers who conducted many interviews, we expect there to be some, but not a lot of nonresponses. To find unusual proportions of item-nonresponse, we need to zoom in on interviewers who conducted a fair number of interviews, and who have either a very low or a high proportion of nonresponse.

Figure 3.4 gives an overview of outliers in proportion of nonresponse, both for the complete sample and for the subset of interviewers who conducted 10 interviews or more. For more details, see table in the annex folder with the proportion of item nonresponse per interviewer on items in Module H ("Prop item nonresponse module H.csv"). We particularly recommend looking at interviewers who have conducted 10 or more interviews and had a proportion of nonresponse larger than 50%. Very little or no nonresponse might also be unusual and worthy of our attention. Those interviewers might be investigated more, but should not be flagged solely on this data.

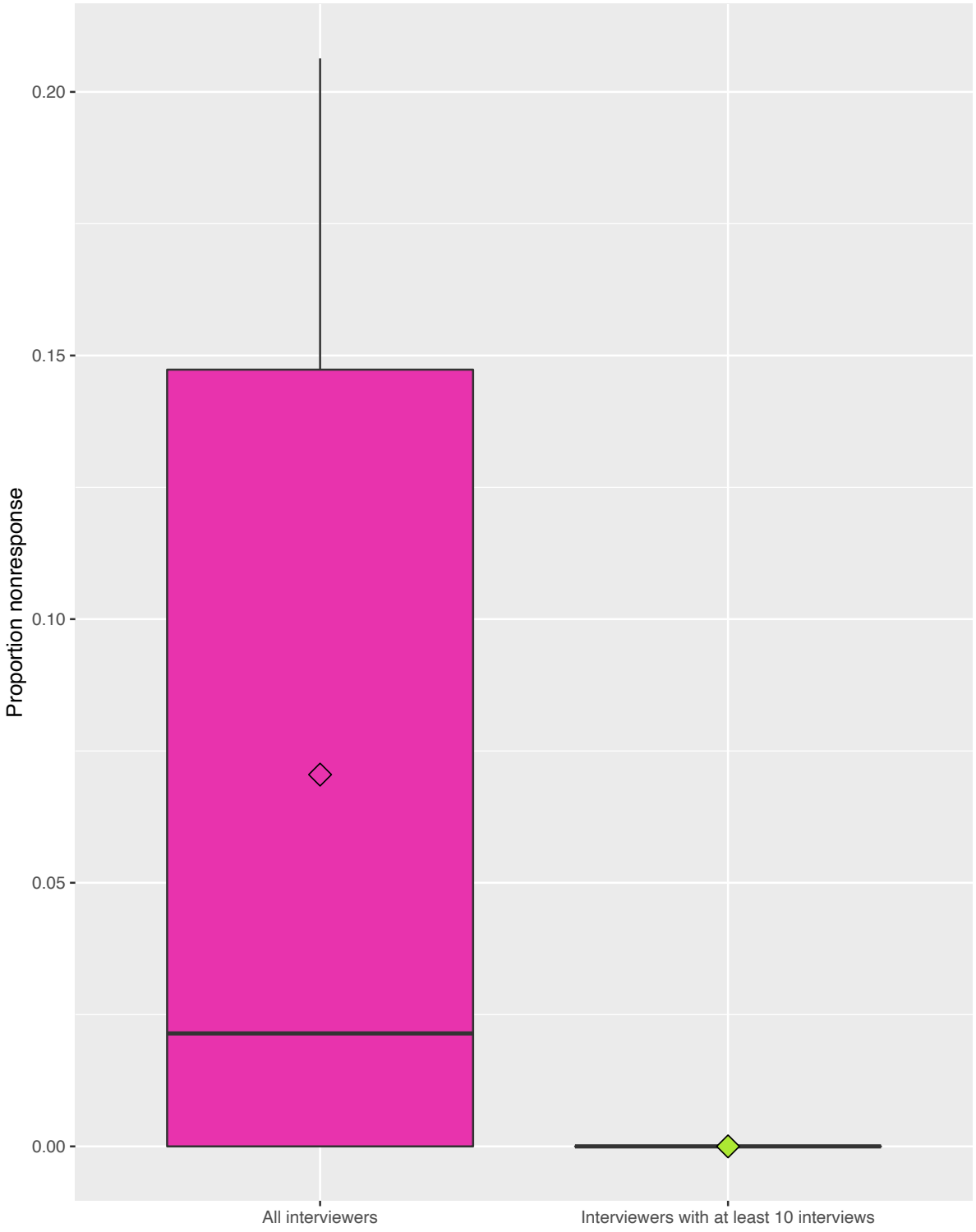


Figure 3.4 Mean proportion of item-nonresponse on items in Module H with labelled outliers
Note: The squares represent the mean of the distribution. Interviewers that are > 1.5*IQR from the borders of the box are defined as outliers.

4 Observed variance between answers of respondents

4.1 Variance between different interviews: Near duplicates

Near duplicates refers to the phenomenon of having two or more interviews with almost identical answers to each survey question resulting in an almost perfect corresponding number sequence between two or more completed interviews. The maximum percentage of items on which a respondent's data matches the data of any other respondent in the sample can be calculated to detect near duplicates cases. It can indicate issues in the interviewing quality (Kuriakose & Robbins, 2016), but also other issues such as data entry errors.

Furthermore, high match rates can also be explained by natural survey features (Simmons, Mercer, Schwarzer, & Kennedy, 2016.) However, near duplicates or high match rates tend to be quite rare, and even more so if one or more near duplicates happen to occur within the interview set of a single interviewer. The same argument applies (albeit it is less rare) to near duplicate sequences on the module level.

High match rates within an interviewer could, therefore, be viewed as an indicator of possible issues in the interviewing process (e.g., related to the interaction of interviewer and respondent, interviewer related, or programming and coding errors). Our analysis strategy is based on the approach used by Kuriakose and Robbins (2016). This means that to avoid overestimating the match rate, we have removed variables with missing data for more than 10 per cent of respondents as well as removed any observation where 25 per cent or more of variables were missing. We qualify match rates of 85% or more between interviews as near duplicates and deem them highly unlikely.

Figure 4.1 shows the distribution of the match rate as percentage of items in the questionnaire with duplicate answers within the same interviewer. For more details, see table in the annex folder with the match rates per interviewer ("Match rates for near duplicates.csv").

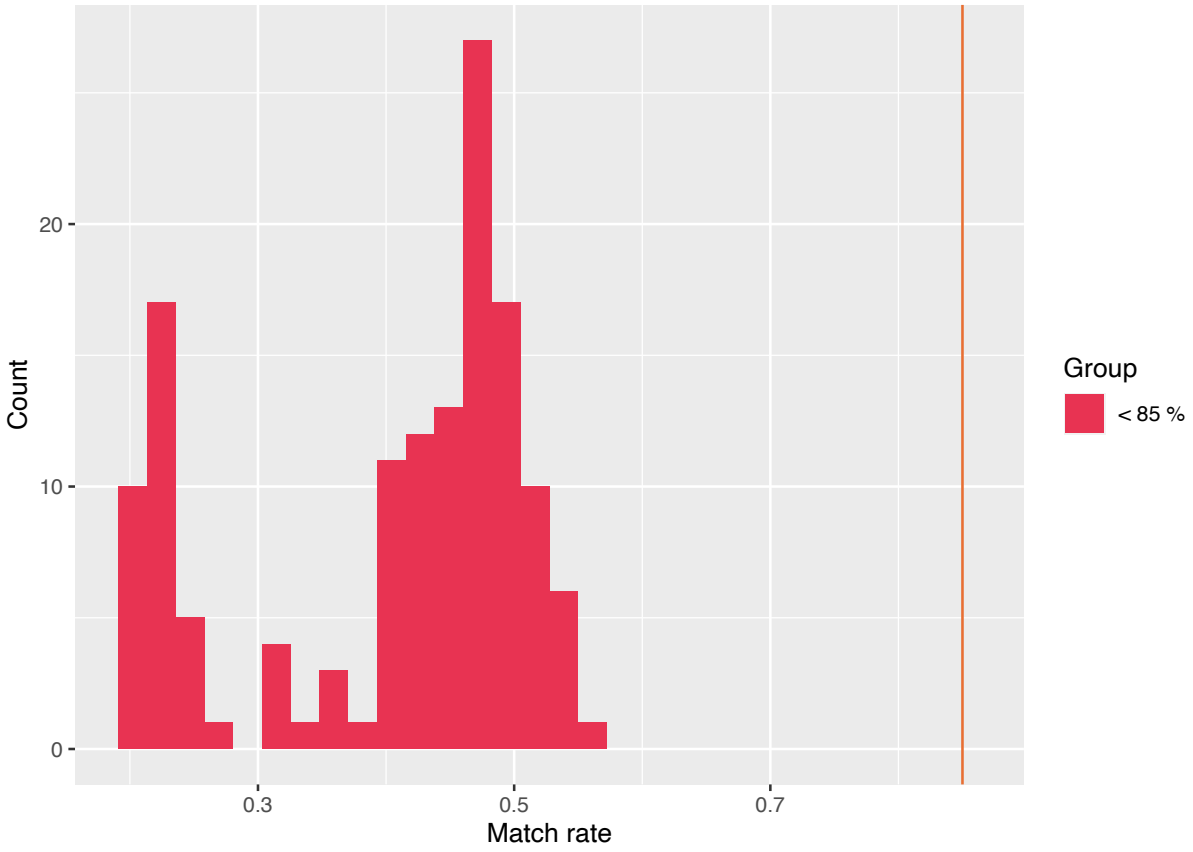


Figure 4.1 Percent Match Rates within interviewers
Note: A match rate of >= 85 % is defined as unlikely. The red line indicates this threshold. Unlikely match rates are displayed in a different colour in the figure.

4.2 Low observed variance within interviews: Non-differentiation

Non-differentiation refers to the tendency to provide the same answer to items in a block of questions (Loosveldt & Beullens, 2017). This response behaviour can be explained by satisficing by the respondents (Chang & Krosnick, 2009). However, Loosveldt and Beullens (2017) were able to observe interviewer effects on the respondent's tendency to choose a response category that is the same as the response category for the previous item. These interviewer effects can, for example, be the result of the absence of challenging the respondent's satisficing tendency by the interviewer, leading the respondents, or completing the questions on behalf of the respondent. This type of undesirable interviewer behaviour can be investigated by looking at the variation in non-differentiation levels between the sets of interviews conducted by a single interviewer.

In our analysis approach, we measure non-differentiation by calculating the Mulligan Score (Mulligan, Krosnick, Smith, Green, & Bizer, 2001). It measures the average mean square root of the absolute difference of all answers within an item block for each respondent. For easier interpretation, we have normalized the reference scale and reversed the score, so that 0 indicates the least non-differentiation and 1 indicates the most non-differentiation. We define a Mulligan Score 1.5 times out of the Interquartilerange (IQR) as improbable.

Non-differentiation results for Module H

Figure 4.2 shows the average Mulligan Score for each interview and interviewers with at least four interviews for module H. We made this differentiation, because the probability that outliers are incidental findings is lower for interviewers with at least four interviews. Interviewers that are defined as outliers are labelled in the figure. For more details, see table in the annex folder with the average Mulligan Score for non-differentiation in module H per interviewer ("Non-differentiation in module H per interviewer.csv").

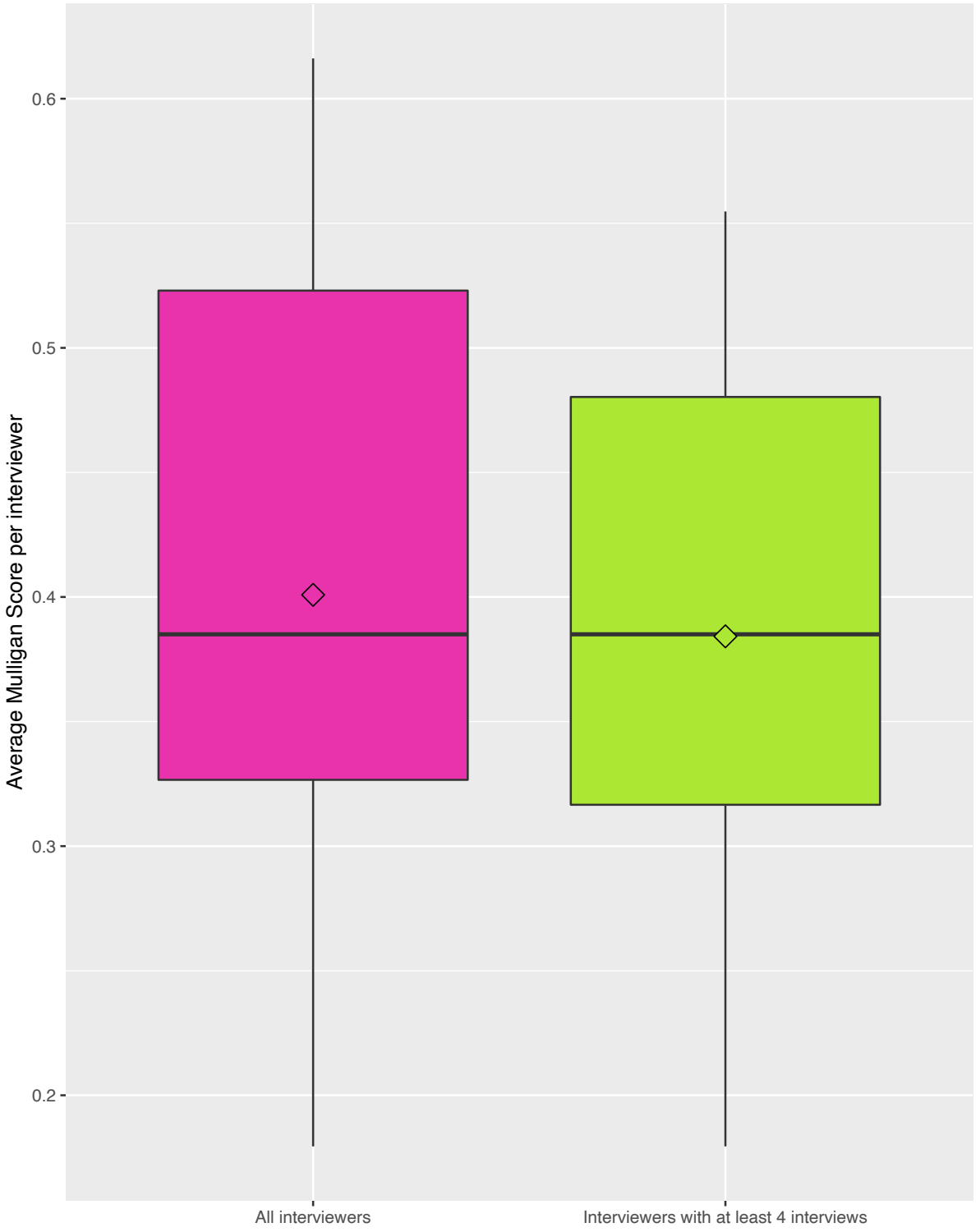


Figure 4.2 Mulligan scores for Module H per interviewer

Note: The nondifferentiation index (Mulligan Score) ranges from 0 (indicating the least non-differentiation) to 1 (indicating the most differentiation). Interviewers that are > 1.5*IQR from the borders of the box are labelled.

4.3 Variance within and between interviews: Difference in latent variables

In this section, we investigate the impact of interviewers on latent variables from item batteries in the questionnaire. We look at the answer patterns to set of items within an interview and compare it to the overall dimension built by the answers pattern the other respondents. The extent to which the differences of the answers of respondents is related to allocation within an interviewer can ascertain the effect interviewers have on answers and can help identify issues regarding interviewer behaviour.

For this purpose, we run a categorical principal component analysis (CatPCA) over a selected battery of items. The categorical principal component analysis is conducted on a set of categorical variables. It allows analysing the relationship between multiple variables while reducing the dimensionality of the data to facilitate interpretation. The items selected for the analysis correspond to the battery of items on trust in institutions from section B of the ESS core questionnaire (trstprl, trstlgl, trstplc, trstplt, trstprt, trstep, and trstun) for which we can expect a dimension ranging from 'no trust at all' to 'complete trust.'

To estimate the effect that interviewers have on the answers of respondents, we calculate (a) the mean of component scores of the first component across interviews conducted by the same interviewer and (b) the standard deviations of those scores. We limit the analysis to interviewers with at least 15 interviews completed, as suggested by the literature (Blasius & Thiessen, 2018), if the maximum number of interviews per interviewer in the dataset is greater than 15. Otherwise, we limit the analysis to interviewers with at least 10 interviews.

A. Mean of Component Scores (between-interviewers variance)

The component scores indicate the distance of each case from to the overall estimated dimension (centered at 0). By calculating the mean of the component scores per interviewer, we can estimate how strongly the answers of respondents interviewed by each interviewer related to the overall component. Figure 4.3, shows the distribution of the mean of the component scores per interviewer. Means that are very different to zero indicate that respondents allocated to those interviewers answered very differently to the battery on trust in institutions compared to the whole sample. For example, a very high mean would indicate that respondents from interviewer X tend to show much more trust in institutions compared to the overall trust of respondents in the sample. To facilitate the overview of results, we define and present outliers with ± 2 standard deviations from the interviewers' mean of means.

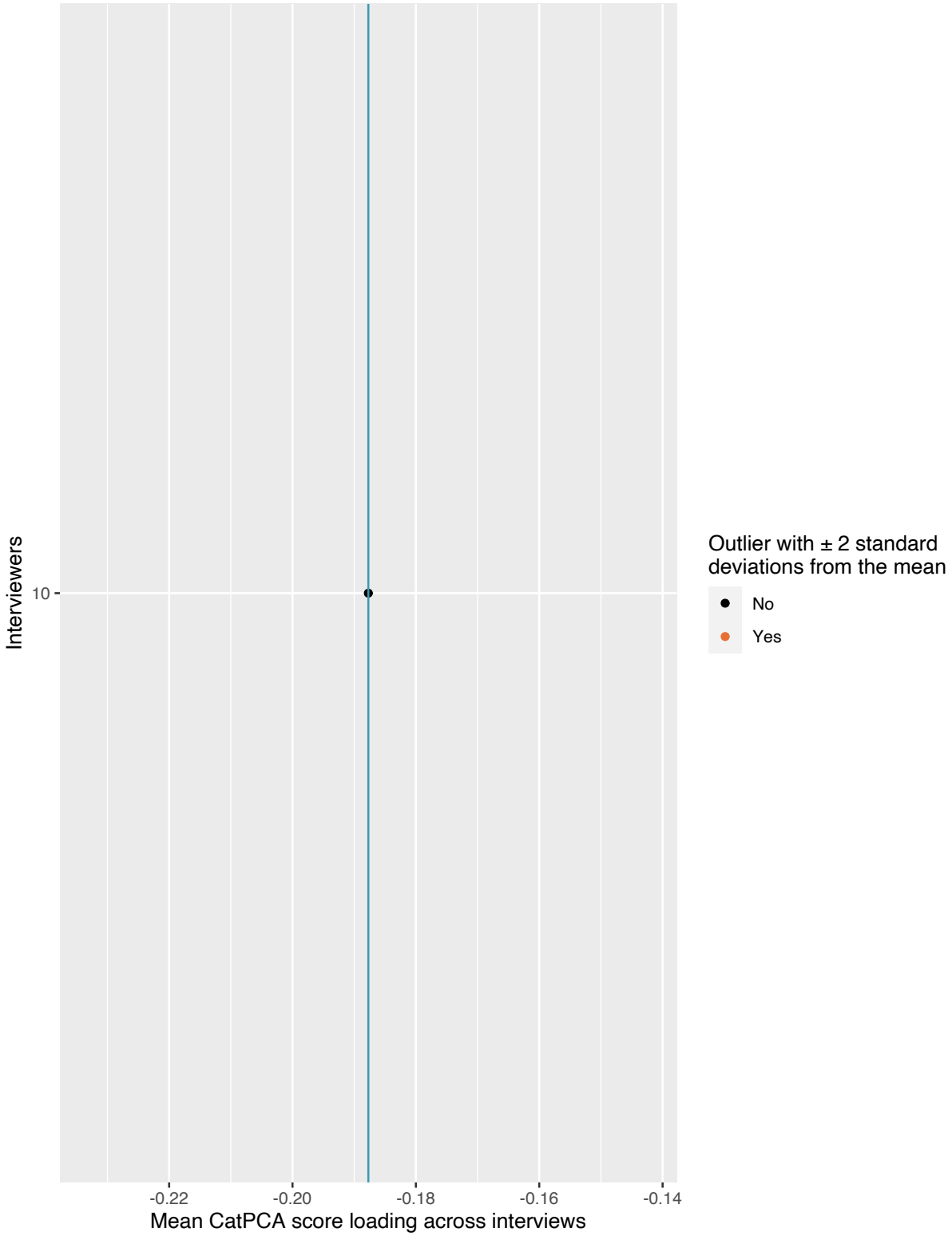


Figure 4.3 Mean of component scores across interviews

Note: Interviewers whose mean score exceeds ± 2 standard deviations from the mean across interviewers are highlighted.

B. Standard Deviation of Component Scores (within-interviewer variance)

It is also important to observe how different or similar the answers of respondents interviewed by the same interviewer are. By calculating the standard deviation of the component scores per interviewer, we can estimate the similarity of the answers of respondents interviewed by one interviewer. Figure 4.4 shows the distributions of the standard deviation of the average scores from the interviews within each interviewer. Large standard deviations indicate that answers of respondents interviewed by the same interviewer are very different with regards to the principal component. In contrast, small standard deviations indicate that answers of respondents were very similar. For example, very small standard deviations of component scores would mean that most respondents from interviewer X shared very similar trust in institutions compared to views of the whole sample. To facilitate the overview, we define and present outliers with ± 2 standard deviations from the interviewers' mean of standard deviations.

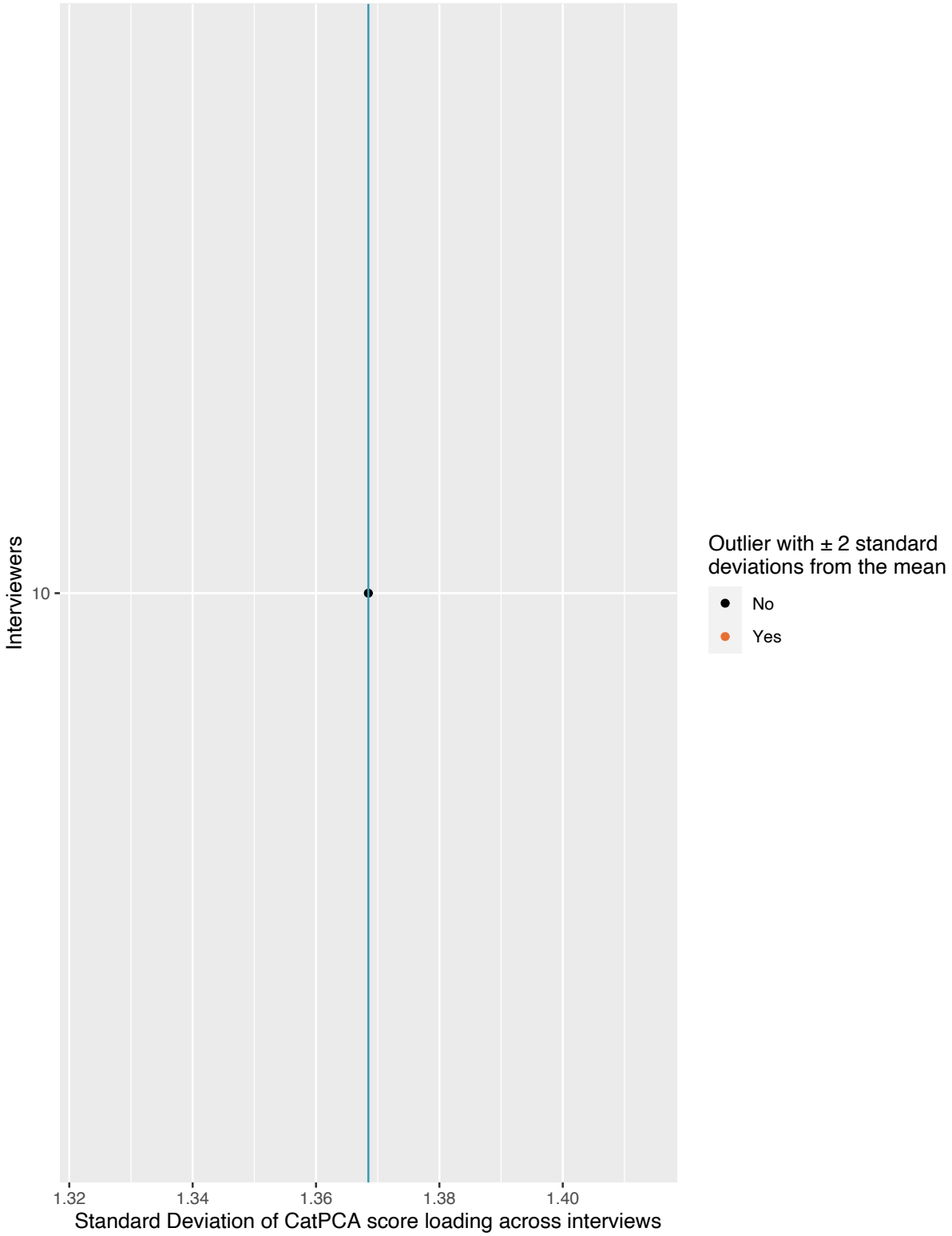


Figure 4.4 Standard deviation of component scores across interviews Note: Interviewers whose mean score exceeds ± 2 standard deviations from the mean standard deviation across interviewers are highlighted.

When interpreting these results, we recommend looking at both the means and standard deviation of component scores to get a better picture of the answer patterns of respondents within a specific interviewer. For more details, please see the table in the annexe folder with means and standard deviation of the component scores ("Latent variable.csv").

Please note that these indicators have been applied by researchers to detect highly unlikely answer patterns in the data being produced by a single interviewer (Blasius & Thiessen, 2018). However, the analysis does not allow us to derive the causes of outliers (as it is the case for most indicators in this report). It is necessary to conduct further investigation on flagged interviewer in the field in order to explain the reasons for any observed issues. It is highly encouraged to conduct back-check and close monitoring of flagged interviewers.

References

Blasius, J., & Thiessen, V. (2018). Perceived corruption, trust, and interviewer behavior in 26 european countries. *Sociological Methods & Research*, 0049124118782554.

Chang, L., & Krosnick, J. A. (2009). National Surveys Via Rdd Telephone Interviewing Versus the Internet: Comparing Sample Representativeness and Response Quality. *Public Opinion Quarterly*, *73*(4), 641–678. https://doi.org/10.1093/poq/nfp075

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

Loosveldt, G., & Beullens, K. (2017). Interviewer effects on non-differentiation and straightlining in the european social survey. *Journal of Official Statistics*, *33*(2), 409–426.

Mulligan, K., Krosnick, J. A., Smith, W., Green, M., & Bizer, G. (2001). Nondifferentiation on attitude rating scales: A test of survey satisficing theory. *Unpublished Manuscript*.

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

Stoop, I., Briceno-Rosas, R., Koch, A., & Vandenplas, C. (2018). *Data falsification in the european social survey?* European Social Survey.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883.

Vandenplas, C., Beullens, K., & Loosveldt, G. (2019). Linking interview speed and interviewer effects on target variables in face-to-face surveys. *Survey Research Methods*, *13*(3), 249–265.