ETC5543 Business Analytics Creative Activity

Project Report

Department of Econometrics and Business Statistics

Monash University

-

The Australian Comparative Study of Survey Method Data Quality Project.

-

Xianghe Xu

32929447

Project Github Repository Link

30 October, 2023

Abstract

This report conducts a comprehensive examination of survey response data quality through a comparison of probability-based methods with non-probability-based methods, using the ACSSM dataset. Various relevant survey response metrics are applied to the dataset variables, enabling the assessment and comparison of response data quality between probability-based and non-probability-based survey methods, particularly regarding the number of careless survey responses in each survey. The results of the comparative analysis demonstrate that probability-based methods exhibit fewer careless responses compared to non-probability methods across nearly all survey response metrics used in this project. This finding supports the hypothesis that non-probability-based methods result in poorer data quality.

This report is for ETC5543 Business Analytics Creative Activity internship project by student **Xianghe Xu**, the github repository for this project is https://github.com/xxuu0086/ETC5543_Internship_Project .

Contents

Introduction and Motivation

In the age of big data, survey research stands out as a fundamental tool for gathering valuable insights from the population. The quality of the responses collected holds paramount importance in making informed decisions. Recent advancements in technology and communication methods have led to a rise in surveys that accessible exclusively to specific groups through online platforms and social media. These surveys are designed for particular purposes, offering less costly and time-efficient data collection. This approach is commonly known as non-probability-based survey methods, in contrast to surveys that aim to collect data from the entire population, referred to as probability-based methods.

Probability-based and non-probability-based survey methods represent two distinct approaches to data collection, prompting our curiosity about the data quality associated with each. However, response quality in non-probability online panels has received limited attention, and there is a scarcity of research on the performance of non-probability-based methods. The study in this project sets out to address the fundamental question: How do probability-based and non-probability-based survey methods compare in terms of data quality?

To investigate this question, the Social Research Centre has conducted a study in 2022 known as the Australian Comparative Study of Survey Method (ACSSM). This study is the first of its kind in Australia, delving into the examination of probability-based and non-probability-based survey methods. The primary aim of this study is to evaluate contemporary and emerging practices for general population surveys. It involves using different survey methods and approaches to collect survey data for the ACSSM study.

In this project, survey data will be drawn from the ACSSM study. The principal aim of this project is to employ a variety of survey response metrics, including individual assessments of metrics such as speeding, straight-lining, outlier analysis, and more. The objective is to conduct a comprehensive comparison of survey response quality between probability-based and non-probability-based survey panels. This comprehensive evaluation will encompass assessments of response completeness, accuracy, reliability, and an overall comparison within both probability-based and non-probability-based surveys.

By consolidating the results from each metric and aggregating the outcomes of all metrics, we will draw a final conclusion regarding the comparison of data quality between probability-based and non-probability-based methods. The ultimate goal is to find evidence that support the hypothesis suggesting that non-probability-based panels exhibit worse data quality when compared to probability-based panels.

Survey response metrics

Surveys are fundamental tools for collecting data, whether from the entire population or a specific group, for experimental or non-experimental research. In real life, we've all experienced or completed surveys in various aspects of our lives, such as student surveys or experience surveys etc... There are numerous incentives that motivate people to participate in surveys. Some do it to contribute to research or a study, while others may be requested by organizations. Some participate for a chance to win prizes or receive cashback after completing the survey. These reasons drive individuals to complete surveys. However, this poses a challenge as some respondents may rush through surveys, investing minimum amount of attention and effort.

This issue is particularly relevant for self-report data collections, commonly used in both experimental and non-experimental psychology, often gathered through online measures. Invalid data can be present in such collections for various reasons, with one major factor being careless or insufficient effort (C/IE) in responding (Curran, 2016). In this project, the terms "careless responding" and "insufficient effort responding" will be used interchangeably. They have the same meaning, which implies that the survey responses provided by the respondents lack sufficient effort and attention when completing the surveys.

The primary objective of this project is to compare the data quality of surveys between probability-based and non-probability-based methods. This raises the question of what constitutes data quality and how it can be measured across different surveys. Survey response metrics provide the answer. To measure data quality,

the first step is to use these survey response metrics to identify careless or insufficient effort responses in each survey method. A survey with a low number of careless responses indicates that most participants made a genuine effort to answer the questions carefully, implying high data quality. Conversely, a survey with a high number of careless responses suggests that participants did not pay much attention and may have randomly answered questions, indicating lower data quality. Therefore, the number of careless responses in a survey becomes an important indicator for comparing data quality across different survey methods.

To initiate this project, it's essential for me to gain a comprehensive understanding of the definitions and practical applications of the available survey response metrics. This entails not only comprehending the significance of each metric but also mastering the techniques for their implementation. My objective is to skillfully apply these metrics to the ACSSM dataset, as well as various variables within it, in order to identify careless responses. Simultaneously, I need to remain aware of the limitations associated with these metrics.

In my journey to learn and master these metrics, I've taken the opportunity to review several articles dedicated to survey response quality. These articles have provided valuable insights into the key metrics that are essential for our project. I have summarized some of the key metrics as follows:

Metrics Descriptions

Response time analysis

- Response time analysis, often referred to as speeding, is a metric that focuses on the time taken to complete a survey or answer questions in a section of a survey. This metric typically relies on page time, which is the duration between the initiation and submission of each survey page or section online. Specifically, it examines instances of extremely small page time values, as such values indicate minimal cognitive processing of survey content. Moreover, this metric involves the use of a "cutoff" time, which is a manually set threshold. Respondents whose page time falls below this "cutoff" time are considered to be responding "carelessly."
- All the surveys in the ACSSM study have recorded the time taken by each individual to complete each section of the survey, so response time analysis can be applied to the ACSSM dataset.

Straight lining

- Straight lining, also known as response pattern analysis, refers to the tendency of respondents to select the same or very similar answer options for each item in a grid. When respondents consistently provide identical responses to a set of questions, it creates a pattern that resembles a straight line, hence the name "straight lining" for this metric. This uniformity in responses across a set of questions suggests that the survey respondents are making minimal effort when completing the survey. In practice, the straight lining are likely to be found for a set of questions with similar wording and the same answer scale. I will be looking for such questions in the ACSSM dataset.

Consistency approach

- This metric goes by various names, including individual consistency, inconsistent approach, psychometric antonyms, and individual reliability, and other names. Despite the different terminology, these terms all refer to the same concept, which involves assessing the consistency of survey responses provided by an individual. In essence, this approach typically involves using matched item pairs and comparing the responses to one item with the responses to another item. The matched item pairs should exhibit either a high positive correlation or a high negative correlation. Consistency in the paired responses is essential to ensure that survey respondents are providing thoughtful and careful answers. In contrast, inconsistent paired responses would suggest that respondents are answering the survey carelessly and randomly.

Outlier analysis

- Outlier analysis is a metric that focuses on data points that are unusual relative to the remainder of a distribution, often seen as extremely values in a distribution. This typically occurs when individual survey respondents who are responding without sufficient effort differ from their more thoughtful counterparts.

Item non-response

- Item non-response or missing data in the dataset indicates that survey respondents are attempting to skip survey questions or provide non-substantive answers. This is often observed as empty answers or responses like "don't know" or "don't want to say." Respondents who tend not to provide much effort in thinking about a correct answer to a complex question are more likely to skip the question, which can be considered as "careless" respondents.

Midpoint selection

- Midpoint selection is a response metric that refers to respondents selecting the middle answer from a list of answer options. The middle answer on the scale is often "Neither agree nor disagree" or "neutral". Similar to the "item non-response" metric, the behavior of choosing the middle value answer can sometimes be interpreted as respondents who are not willing to put in much effort to consider their true response to a question. Instead, they pick for the middle point, indicating neither agreement nor disagreement with the survey question. In this metric, I will focus on questions with ordinal responses, such as "Strongly agree", "Agree", "Neither agree nor disagree", "Disagree", and "Strongly disagree." These types of questions are more likely to elicit middle point selections, which can be considered a form of insufficient effort responding.

Bogus Items

- Bogus items are questions that have an obvious correct answer, making an incorrect response indicative of inattention to the question. If a survey respondent fails to provide the correct answer to a question with an obvious solution, it implies that the respondent is responding to the survey carelessly, without properly reading the question. This metric is straightforward to implement; it involves identifying bogus items in the survey data and examining responses that deviate from the correct answer.

Mahalanobis Distance

- Mahalanobis distance is a metric similar to outlier analysis. Both metrics seek extreme values that deviate from the main distribution. The key difference is that Mahalanobis distance is a multivariate outlier technique, whereas outlier analysis examines single variables. This metric calculates the distance of data points from the multivariate center, allowing it to determine whether a respondent falls at the edge of the multivariate distribution formed by responses to all items. If a respondent is positioned far from the distribution in multivariate space, they may be considered unusual and potentially careless in responding their surveys.

The descriptions above provide a brief overview of the metrics used in this project. It's important to note that there are quite many survey response metrics available, and not all of them can be applied to the ACSSM dataset in this project. Only certain metrics are suitable for this dataset, so I won't list descriptions for all of them here. Some metrics can be used multiple times within the dataset, like the consistency approach, which allows for checking the consistency of more than one pair of responses in the dataset.

With these metrics now at the forefront of our considerations, we can proceed to conduct a thorough examination of the ACSSM dataset.

Data Description

After familiarising ourselves with some of the popular survey response metrics discussed above, we have gained an understanding of their meaning and practical implementation. Now, it's time to delve into the dataset and establish connections between these metrics and the data. The data set used in this project is sourced from the study 'The Australian Comparative Study of Survey Method'. This study encompasses survey data from eight surveys, with four of them being probability based and the remaining four non-probability based. The data set comprises 6,043 rows of observations, indicating the participation of 6,043 survey respondents in the study. It has 290 columns, implying presence of 290 variables within the data set. It's important to note that not all 290 variables are survey questions; some variables represent the survey questions posed to each respondent and record their responses. Other variables contain additional information about the surveys, such as the respondent's 'serial number', 'ID', and 'DataSource', which records additional information about the survey respondents. As discussed earlier, there are many response metrics available and not all metrics will be used in this data set, similarly, not all variables will be used to compare the quality of survey data. Therefore, before applying the metrics to the data, I must manually review each variable to identify the right ones for the metrics. Some metrics are straightforward to apply to the variable. For instance, the response time analysis metric only requires the page time. In such cases, I can directly access the variable that contains the page time which is the time taken to answer the survey questions and use it for response time analysis, thus determining whether respondents are responding quickly or not. However, some metrics are more complex to implement. For example, the consistency approach requires paired variables that exhibit high positive or negative correlations. Pairing these variables is a manual process.

Now, let's take a quick glance at the data set to get an initial impression of its structure and determine whether any cleaning and tidying are required.

Table 1: Top 10 variables with a large number of missing values.

variable	n_miss	pct_miss
RR1	6043	100.00000
RR2	6043	100.00000
LOTE	6043	100.00000
MOBAPPT	6043	100.00000
LINA_VALI_REAPPT	6043	100.00000
LINA_VALI_REAPPT_LINK	6043	100.00000
DEVICE_SWITCH	6043	100.00000
DEVICE_TYPE2	6043	100.00000
LINA_VALI_PREINTRO2	6022	99.65249
AGE_Codes	6020	99.61939

Table 1 displays the top 10 variables with the greatest number of missing values. In this table, the left columns show the names of the variables in the data set, while the columns 'n_miss' and 'pct_miss' represent the number of missing values and the percentage of missing values within each corresponding variable. It is evident that each of the top 10 variables has almost 100% missing values. A variable with a high percentage of missing values is not useful for analysis and should be ignored or removed from the data set. In fact, variables with more than 10% missing values should not be used in any analysis, especially in data sets with a large number of observations, as a high percentage of missing values can introduce bias into the analysis.

There are 91 variables in this data set with more than 10% missing values. These variables will be removed from the data set to ensure that the analysis in this project is based on variables with relatively completed data. This step will contribute to more accurate and less biased results in the survey method data quality analysis.

The variables with more than 10% missing values have been successfully removed from the dataset. As previously mentioned, the ACSSM dataset originally contained data from 8 surveys. However, I was instructed

by the social research center not to use two of them. As a result, I am left with data from 6 surveys that can be used for comparing survey data quality. Out of these 6 surveys, two are probability-based, and four are non-probability-based. While the number of probability-based surveys is now less than the number of non-probability surveys, this discrepancy does not pose a problem because the goal is to compare the data quality between the surveys.

To facilitate the comparison and improve the clarity of the analysis, I have added two new columns to the dataset. The first column labels whether a respondent is from a probability-based survey or a non-probability-based survey. The second column contains the full names of the surveys, corresponding to the index numbers in the 'DataSource' variable. This is particularly useful because, in the ACSSM data, each survey is represented by digits in the 'DataSource' variable, which can be confusing at times. By adding this new column with corresponding survey names, it becomes easier to identify which survey a respondent belongs to. In this new column, "Life in Australia" and "SMS push to web" surveys are classified as probability-based, while "Panel 1", "Panel 2", "Panel 3", and "Panel 4" surveys are considered non-probability-based.

Now, after cleaning and organizing the data set, the new ACSSM data set contains 4,640 observations and 201 variables.

We have the cleaned ACSSM data set ready. Now, it's time to connect the metrics with the variables in the data set. To simplify the connection process, I have created an Excel spreadsheet. In this spreadsheet, I've listed all the questions from different sections of the survey, along with some potentially applicable metrics listed on the other side of the spreadsheet. Using Excel spreadsheet provides an overview of all survey questions, which will facilitate the connection between metrics and variables. You can find this Excel spreadsheet named "Survey_question.xlsx" in the "Project_related_material" folder.

Analysis

Response Time Analysis

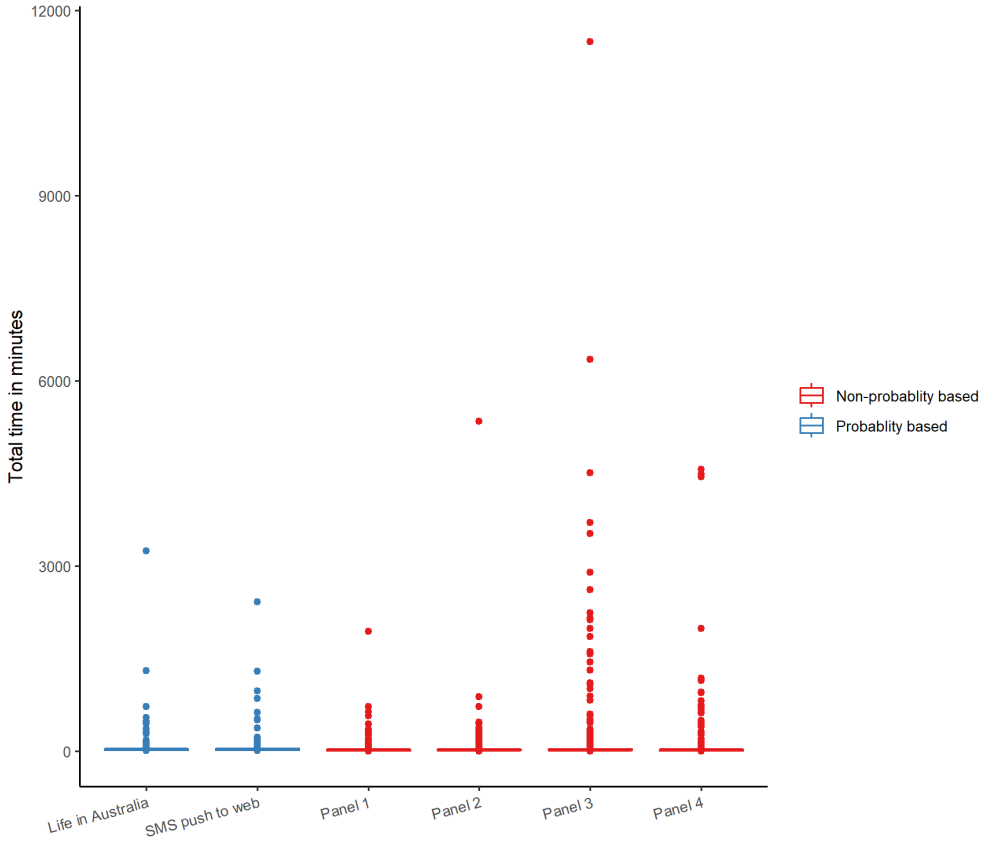

Figure 1: Distributions of time spent answering the survey questions for panels.

Response time, which measures the time it takes for an individual to respond to a set of items, is a widely employed tool for identifying careless and insufficient effort survey respondents(Curran, 2016). We aim to use this response time metric to pinpoint respondents who take an unusually amount of time to answer survey questions. To achieve this, the first step is to gain an overall understanding of the time spent by each survey respondent on the survey questions. I attempted to use a box-plot to illustrate the distribution of time spent for each survey panel.

Above Figure 1 illustrates the distribution of all time spent on answering the same survey questions across different survey panels in the study. From Figure 1, it's clear that the data points are widely dispersed, with many outliers, often referred to as extreme values. These outliers represent respondents who spent an exceptionally long time answering the survey questions. As shown in the distribution plot, it's appeared that a number of survey respondents spent over 3000 minutes, which is equivalent to 50 hours, on completing the survey. This is highly unusual because very few surveys should take more than an hour to complete.

However, labeling those survey respondents who spent an extended amount of time on survey responses as "careless" or displaying "insufficient effort" is a challenging task. For example, consider a scenario in which a survey respondent initially approaches the survey with utmost diligence, paying close attention to each question. However, due to unforeseen circumstances or more pressing tasks, they may need to interrupt the survey and return to it later. In such cases, these respondents cannot be labeled as careless or

showing insufficient effort; they are simply taking longer to complete the survey due to external factors. This demonstrates the complexity of determining whether those who spend an extended time on survey responses should be classified as careless respondents or not.

On the other hand, those who complete a survey in a minimal amount of time can certainly be considered as careless and insufficient effort respondents. Most social research based on survey data relies on the assumption that respondents answer the survey questions to the best of their ability. This, in turn, requires the respondents to meticulously go through all the cognitive steps involved in answering a survey question (Cornesse, 2020). The cognitive response process typically encompasses four essential steps: question comprehension, information retrieval, judgment and estimation, and reporting an answer. Consequently, it's implied that a respondent who responds to a survey with due care and follows these cognitive steps will require a minimum amount of time to thoroughly understand the questions, select the correct responses, and provide answers to the survey.

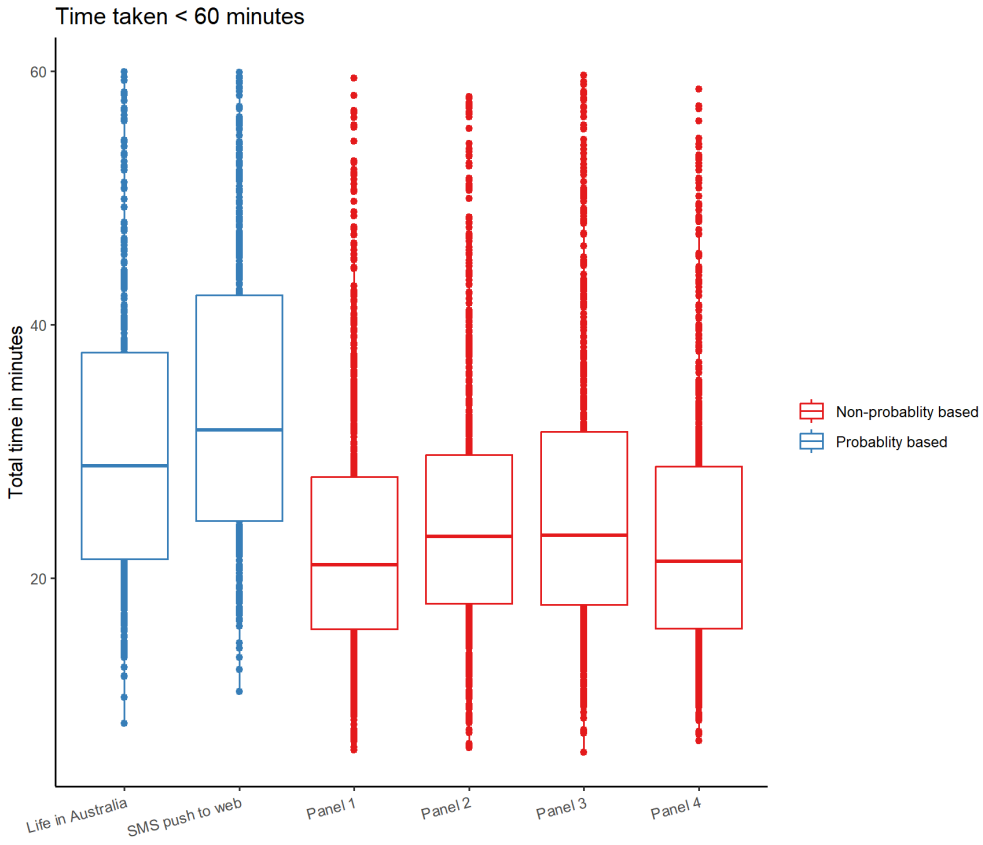

Figure 2: The distribution of time spent answering the survey questions for each panel within an hour.

Since the distribution of completion times for survey questions is widely spread and we are not specifically concerned with extremely large amounts of time spent on surveys, we can set a time range cap and focus on the lower end of the time range. Figure 2 presents a box-plot distribution of the time it takes to answer survey questions in less than 60 minutes.

When comparing the box-plot distributions in Figure 2, we observe that these distributions are quite similar. They share a similar range and have comparable inter-quartile ranges. However, the non-probability-based surveys exhibit lower median values compared to the probability-based surveys. This implies that more respondents from the non-probability-based surveys spend less time answering the survey questions.

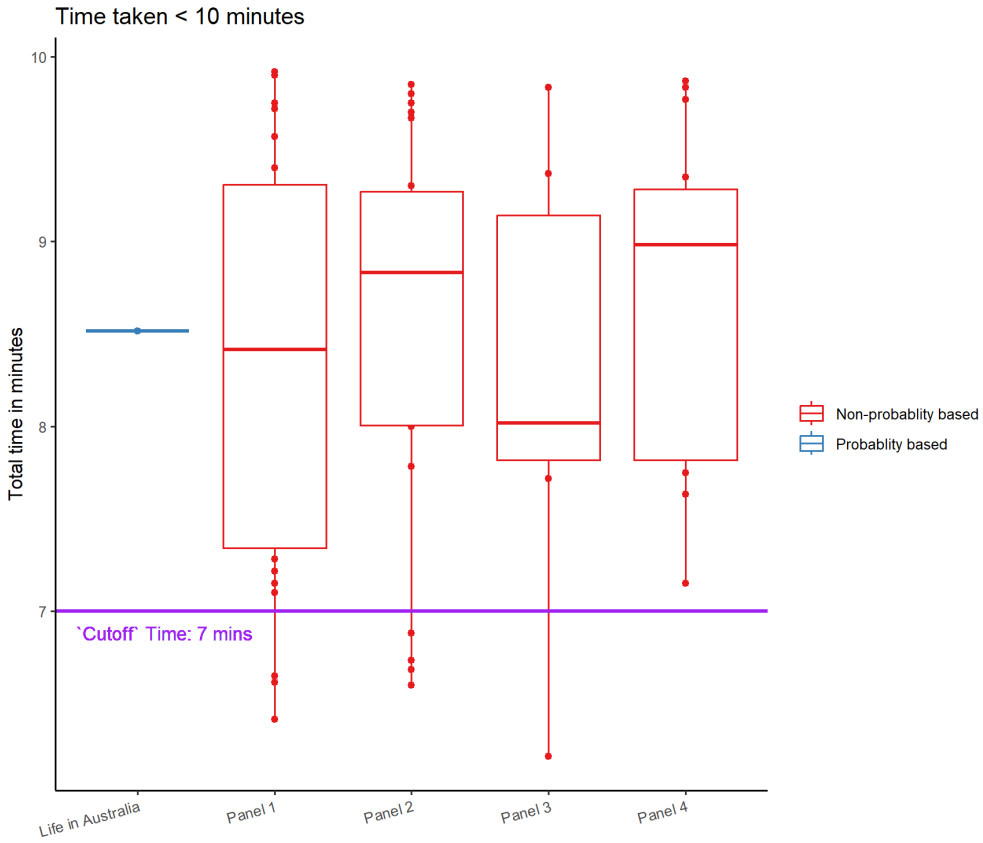

Figure 3: The distribution of time spent answering the survey questions for each panel within 10 minutes.

In Figure 2, we established a time range cap of 60 minutes. Now, we can further narrow down the time range. Figure 3 displays box-plot distributions for each survey with a time spent of less than 10 minutes. This plot also includes a purple cutoff time line at y = 7 minutes, which I manually set. My rationale behind this cutoff line is that I believe each respondent requires an average of 6 seconds to answer each of the 70 questions in a survey. Therefore, to complete all survey questions, a respondent would need at least 7 minutes. Anyone who takes less than 7 minutes to answer all the questions can be considered as engaging in "careless responding".

Upon comparing the box-plot distributions, it becomes evident that many respondents from non-probability based surveys used less than 10 minutes to complete the survey. In contrast, there was only one respondent in the probability-based surveys who took about 8.5 minutes to complete it. Additionally, about eight respondents from the non-probability-based surveys took less than 7 minutes to complete the survey. According to my decision rule of 7 minutes, these eight respondents are considered to be careless and showing insufficient effort in their responses. In contrast, none of the respondents from the probability-based survey completed the survey in less than 7 minutes. Consequently, the presence of eight careless respondents from the non-probability-based survey suggests that people from the non-probability-based surveys tend to rush through the surveys. This implies that non-probability-based surveys may have worse data quality compared to probability-based surveys when considering the response time analysis metric.

Straight lining

As previously mentioned, straight lining happens when a respondent provides identical answers to a set of survey questions, suggesting minimal effort in their responses. In the ACSSM data set, we have identified

four groups of survey questions, each featuring questions with similar wording and the same answer options. Our objective is to pinpoint the careless and insufficient respondents who exhibit straight lining behavior when answering these four question groups.

Since there are four sets of questions where straight-lining is likely to occur, I have established a decision rule. According to this rule, respondents who straight lining at least two sets of questions will be identified as careless and insufficient respondents. Those who straight lining only one set of questions will not be labeled as careless, as they may genuinely be providing the true answers to those questions.

Figure 4: An example of straight lining

The Figure 4 provides an illustration of straight lining, 'Agree' is chosen for all five questions and the answers form a straight line.

Table 2: '1' = Total number of 'Straight lining' respondents

	0	1
Life in Australia	577	8
SMS push to web	595	5
Panel 1	817	39
Panel 2	809	44
Panel 3	853	39
Panel 4	832	22

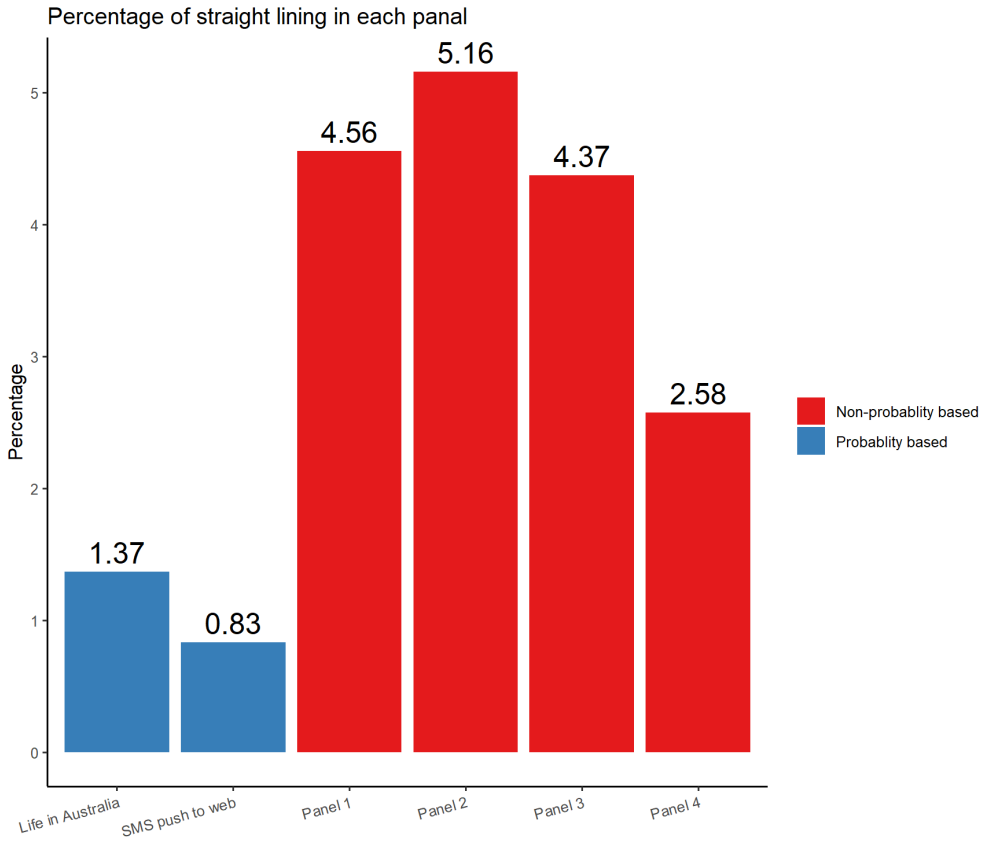

Figure 5: Histogram of the Percentage of Straight Lining in Each Panel.

Table 2 presents the total number of respondents identified as "careless" according to my straight-lining decision rule. This rule categorizes those who attempted straight lining in at least two question sets as careless respondents. In the table, the '0' column indicates the number of respondents in each survey who were not identified as careless due to straight-lining, while the '1' column shows the count of respondents engaging in straight-lining for each survey.

Upon examining the table, it becomes apparent that the probability-based surveys "Life in Australia" and "SMS push to web" have only a small number of respondents in the single digits who engaged in straight-lining. In contrast, the non-probability-based surveys "Panel 1", "Panel 2", "Panel 3", and "Panel 4" each have 39, 44, 39, and 22 respondents, respectively, who were identified as "straight liners". These numbers are more than double the counts observed in the probability-based surveys.

Figure 5 displays a histogram depicting the percentage of survey respondents engaged in straight-lining for each survey. A comparison of the percentages in Figure 5 reveals that non-probability-based survey panels have a notably higher percentage of "straight-lining" respondents in comparison to probability-based survey panels. A higher percentage of "straight-lining" respondents within a survey indicates that a greater number of respondents in that survey are likely to be engaged in straight-lining, while surveys with lower percentages have fewer such respondents.

From both Table 2 and Figure 5, it becomes evident that a higher occurrence of straight-lining is observed in non-probability-based survey panels. This suggests a greater presence of careless respondents in non-probability-based panels and implies that these panels have poorer data quality in terms of straight lining.

Consistency Approach

This metric depends on paired items, which are selected pairs of variables from the data set. In a high-quality survey where respondents pay close attention, the responses to these paired items should demonstrate consistency. However, if there are inconsistencies in the responses, it implies that the respondents may not have given enough attention to reading the questions and providing coherent answers. This situation also raises concerns that respondents might be providing random responses to the survey. When a respondent provides a pair of inconsistent responses, they are categorized as careless or insufficient effort respondents. In this data set, I have identified two pairs of questions suitable for applying this consistency approach.

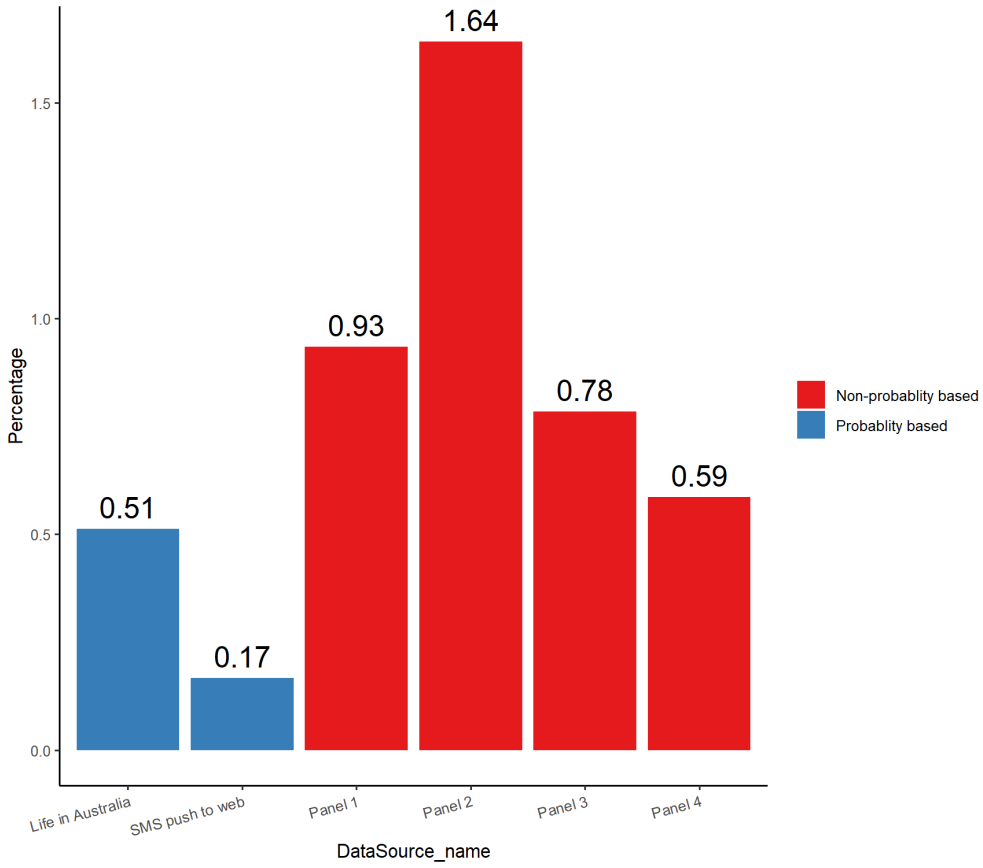

Figure 6: Histogram of the Percentage of inconsistency in responding allowance payment questions

The first pair of variables consists of 'BENTYPE_d' and 'UNPAIDCARE'. 'BENTYPE_d' is a question asking people whether they receive any carer allowance or carer payment from the government. 'UNPAIDCARE' is a question inquiring whether individuals have spent time providing unpaid care in the last two weeks. The instructions for the 'UNPAIDCARE' question specify that even if respondents have received carer allowances from the government, they should select 'Yes' for this unpaid care question. In other words, respondents who have received carer allowances should still answer 'Yes' to the unpaid care question. However, if respondents are not paying close attention and don't read the question instructions carefully, they might mistakenly select 'No' for the unpaid care question because they have received carer allowances from the government.

Figure 6 displays a histogram illustrating the percentage of respondents who provided inconsistent answers to the 'BENTYPE_d' and 'UNPAIDCARE' questions. The histogram clearly shows that a greater percentage of respondents in each non-probability based survey provided inconsistent responses to the 'BENTYPE_d' and 'UNPAIDCARE' questions compared to probability-based surveys. This suggests that a greater presence

of respondents from non-probability based surveys are being careless and not reading the question instructions properly.

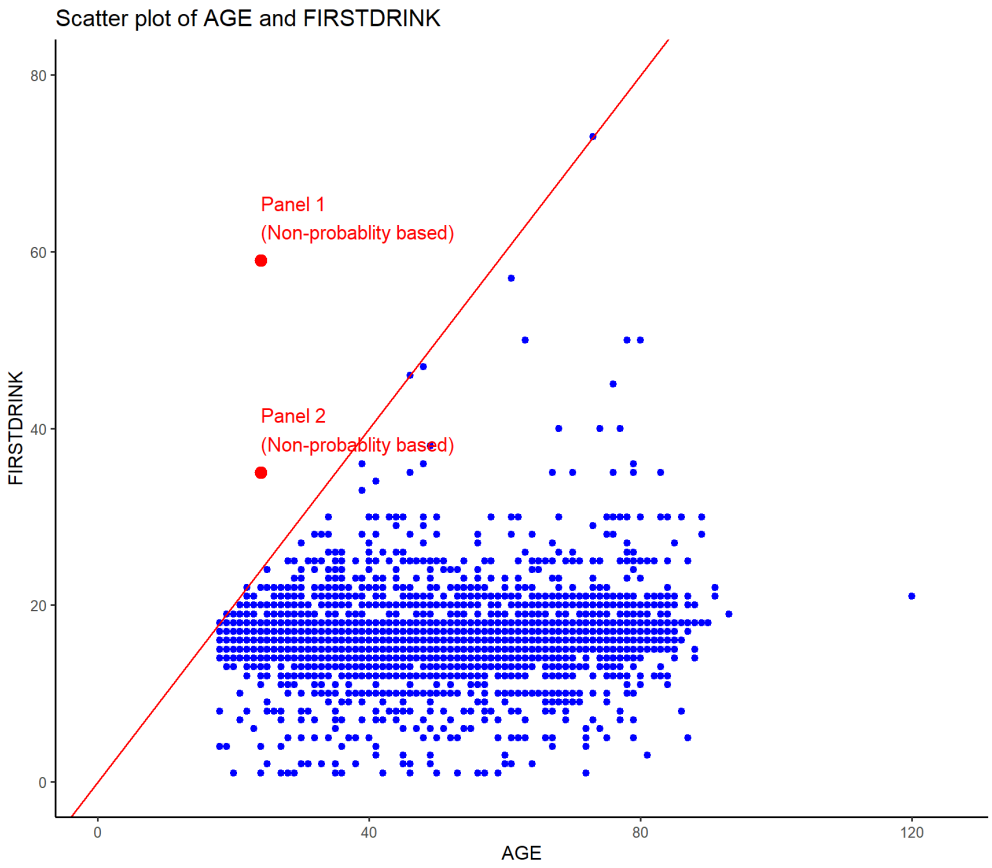

Figure 7: Scatter plot of variables AGE and FIRSTDRINK

The second pair of variables includes 'AGE' and 'FIRSTDRINK'. 'AGE' is a question asking for the respondent's current age, while 'FIRSTDRINK' is a question inquiring about the age at which people had their first alcoholic beverage. Both questions are related to age. It's important to note that I will ignore individuals who have never had their first alcoholic beverage for this analysis. Logically, people could have had their first alcoholic drink in the past or in the current year, but they cannot say they had their first alcoholic beverage in the future. This means that the value of the response for 'AGE' should always be greater than or equal to the value of the response for 'FIRSTDRINK'. Anyone who violates this rule by providing inconsistent answers should be considered a careless respondent.

Figure 7 is a scatter plot of the variables 'AGE' and 'FIRSTDRINK' with a red 45-degree line dividing the plot into two sections. We know that the value of 'AGE' should always be greater than or equal to 'FIRSTDRINK'. If this is the case, all data points should fall on the red line or on the right side of the red line. However, there are two respondents from non-probability based surveys who violate this rule; they appear on the left side of the 45-degree line. This implies that these two respondents are in their age of 20s but they claim to have had their first alcoholic beverage in their 30s and 50s, respectively. This is impossible, suggesting that these two respondents from non-probability based surveys are providing careless responses to the 'AGE' and 'FIRSTDRINK' questions.

Both sets of paired variables, as demonstrated in Figure 6 and Figure 7, reveal a higher presence of careless responses in non-probability based survey panels.

Outlier analysis

Outlier analysis is a metric that focuses on extreme values within a distribution. If a survey respondent's answer deviates significantly from the majority of other responses and this answer is exceptionally unusual, it could arguably be considered a case of careless and insufficient effort in responding.

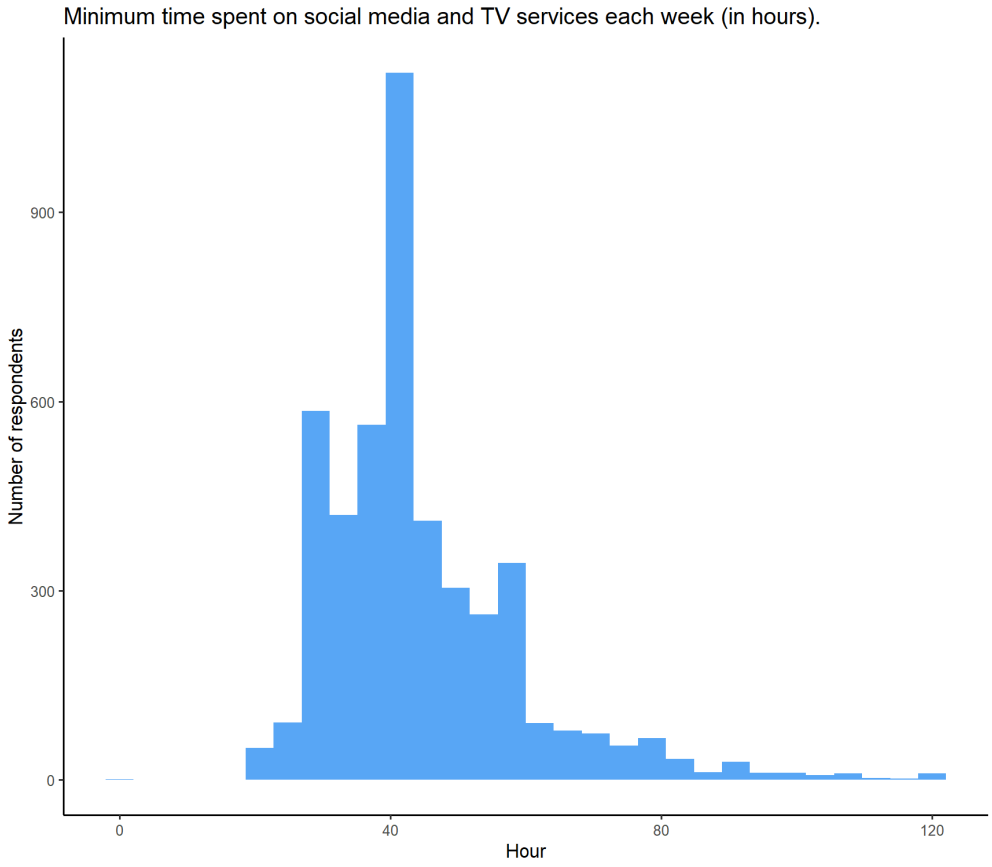

Figure 8: A distribution of the total time spent on video services and TV per week.

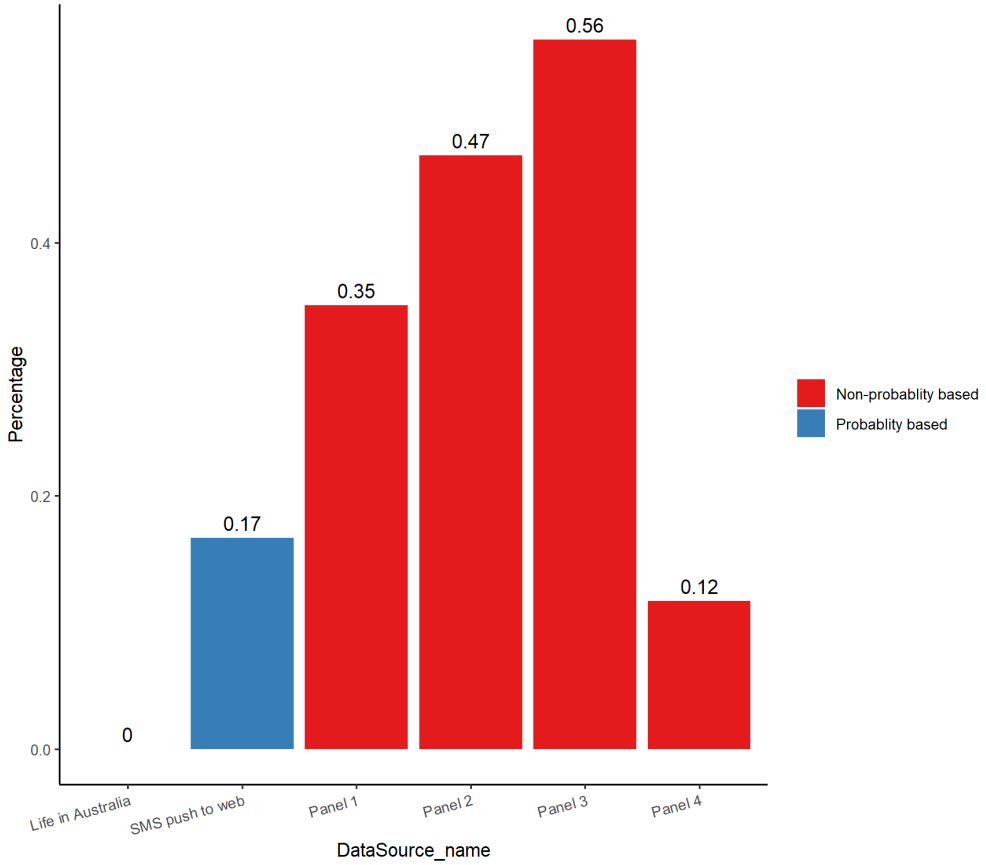

Figure 9: A histogram of the percentage of at least 120 hours spent on video services and TV per week.

In the ACSSM data set, there are six questions that inquire about the time spent on various video and TV services, including 'Free video streaming services', 'Online subscription services', 'Pay TV', 'Free on-demand TV', 'Publicly owned free-to-air TV', and 'Commercial free-to-air TV'. The answer options for these questions are the same, with six possible choices: '0 hours per week', '1-5 hours per week', '6-10 hours per week', '11-15 hours per week', '16-20 hours per week', and 'More than 20 hours per week'. While the hours listed in each answer option may seem normal, even the one that says 'More than 20 hours per week', we might not consider it unusual for someone to spend more than 20 hours watching TV.

However, these six questions pertain to different kinds of video and TV services. To gain a visual overview of the aggregated time spent on all these services, I summed up the hours from all six questions based on respondents' answers. The resulting histogram distribution in Figure 8 displays the total time spent on video and TV services per week. This distribution is right-skewed, with a center around 40 hours per week. The distribution also reveals some extreme values, with a few respondents indicating that they spend over 120 hours on video and TV services, which is quite unusual, given that this would mean spending at least 17 hours each day on such activities.

Figure 9 illustrates the percentages of respondents who report spending at least 120 hours on video and TV services per week in each survey. Allocating 120 hours to watching video and TV can be considered highly unusual, suggesting that these respondents may not have read the questions properly or have provided random answers. In either case, such respondents can be classified as careless respondents. The data in Figure 9 shows that non-probability based survey panels have a majority of careless respondents who report watching more than 120 hours of video and TV per week.

Midpoint selection

Selecting the middle answer from a set of answer options can sometimes indicate insufficient effort in responding, as it suggests that the respondents are not willing to exert the effort needed to choose the correct answer from the provided scale.

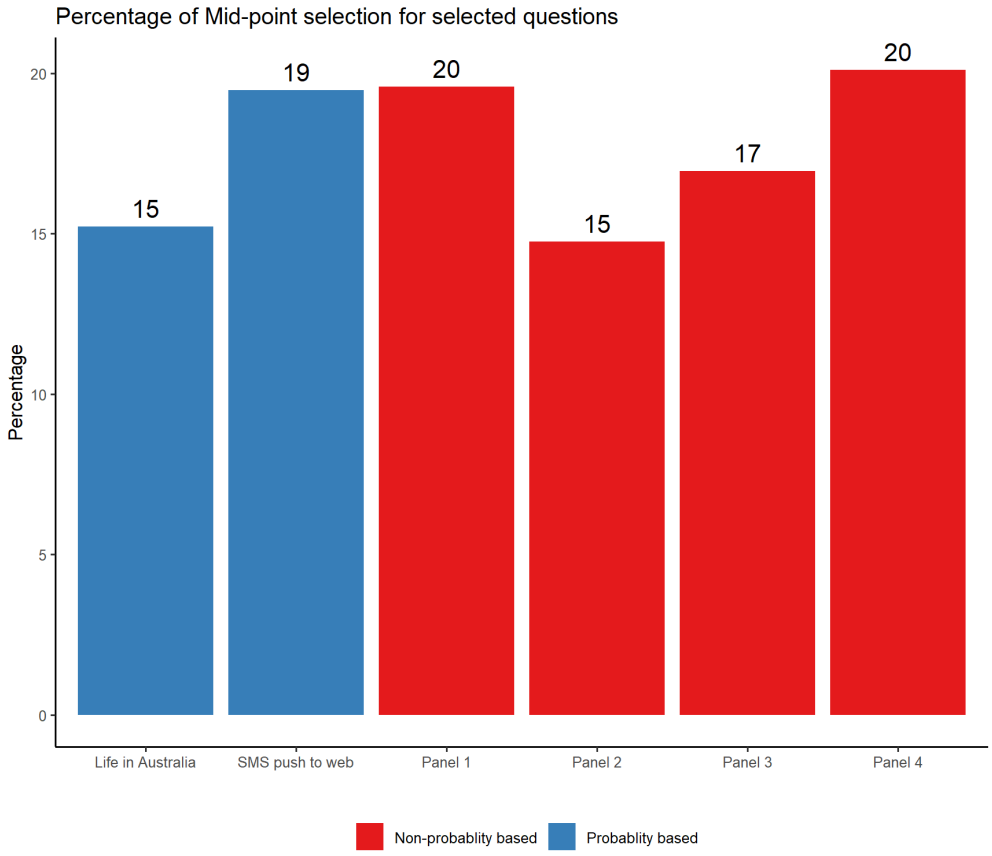

Figure 10: Distribution plot of percentage of selecting middle response for each survey.

Figure 10 presents a histogram showing the distribution of the percentage of respondents who select the middle answer for selected questions across each survey panel. These selected questions have a higher likelihood of triggering mid-point selection, typically featuring ordinal answer scales with options like "Neither agree nor disagree" or "neutral".

The figure 10 indicates that there is a similar percentage of respondents in each survey who choose the middle answer for these questions. This suggests that both the probability based and non-probability based surveys exhibit similar performance in terms of the Mid-point selection metric.

Item non-response

Item non-response refers to respondents choosing 'Don't know' or 'Refused to answer' for survey questions. Individuals who do not wish to answer a question or do not want to expend much effort considering the response are likely to choose 'Don't know' or 'Refused to answer'. The quantity of 'Item non-response' answers in a survey can be indicative of the extent of insufficient effort responding present in that survey.

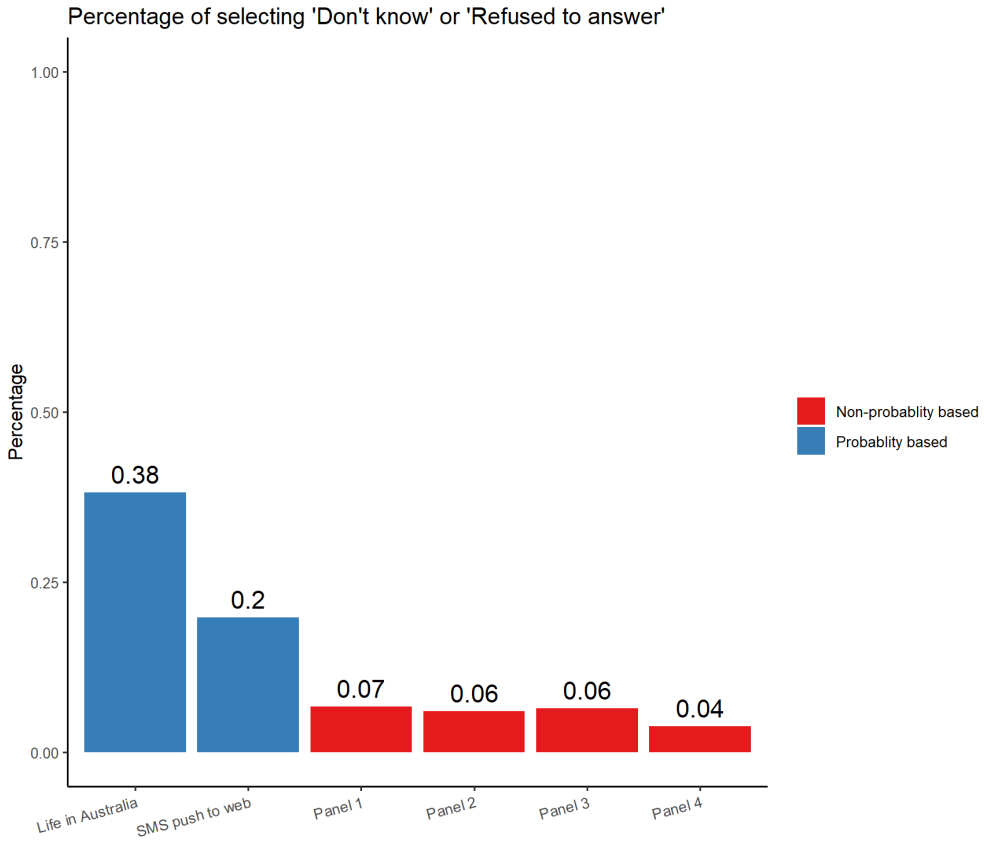

Figure 11: The distribution of the percentage of item non-response in each survey.

Figure 11 illustrates the percentage of respondents in each survey who answered 'Don't know' or 'Refused to answer' for survey questions. The plot reveals that there is a slightly higher percentage of respondents in probability based surveys who provided 'Don't know' or 'Refused to answer' compared to non-probability based surveys. In this scenario, probability based survey panels exhibit a higher prevalence of insufficient effort responding than non-probability based survey panels.

Combine all metrics

Although nearly all the metrics I have used so far indicate that the non-probability based survey panels have worse data quality compared to probability based survey panels in terms of the number of careless or insufficient effort responses in each survey, I would still like to combine all the careless respondents defined by each metric and examine an overall comparison using all the metrics I have employed in this project. It's worth noting that the metrics "Bogus item" and "Mahalanobis distance", which I initially planned to use in this project, will not be included as I couldn't find the right variables for them.

Table 3: Average percentage of careless and insufficient effort respondents

prob_based	Percentage
Non-probablity based	7.12
Probablity based	5.66

Table 3 presents the final result after combining all metrics. On average, non-probability based survey panels

have 7.12% of careless and insufficient effort respondents, which is higher than the 5.66% in probability-based panels in this study. The higher percentage of insufficient effort respondents in non-probability based panels indicates that they have worse data quality compared to probability-based panels.

Chanllenges and Limitation

Chanllenging part

The most challenging task in this project is linking the variables to the metrics. For some metrics, we know the types of variables they require. For instance, response time analysis metrics can only be applied to page time variables, which record the time each individual takes to complete survey questions. So, for this metric, our main task is to find suitable time variables.

However, for some metrics, the process is less straightforward. We must manually identify the variables for these metrics, and this can be a particularly challenging aspect of the project, requiring careful consideration.

Take the consistency metric, for example, which involves working with two pairs of variables. The relationship between these variables isn't always obvious, so we need to explore all possible combinations of responses from the paired survey questions. For instance, some combinations of responses provided by the respondents are abnormal and can be considered as careless and insufficient effort in responding. This means we must thoroughly examine all variables, check all responses for each variable, and search for potential matches for each metric. Often, we need to refer to the data dictionary and the ACSSM study technical report when the meanings of variables are unclear. Additionally, we need to conduct tests on these potential variables to determine their suitability for implementing the metrics.

Limitations

There are several limitations to consider when using response metrics in this project, and these limitations should be kept in mind when presenting the project's results. Some of the key limitations include, but are not limited to:

- **Response time analysis**: Careless and insufficient responding may occur among respondents who spend a long time completing survey questions, but we have ignored them in the analysis because it is challenging to identify them.
- **Outlier analysis**: In the case of outlier analysis, we assume that most responses are provided by respondents who are providing genuine answers, so those extreme values deviated from the majority are unusual.
- **Straight lining**: Respondents are identified as engaging in insufficient responding if they provide the same answers to at least two groups of questions. However, it is possible that they are providing true responses.
- **Mid-Point selection approach**: Respondents may choose the middle answer as it represents the most accurate response for them.

Conclusion

In conclusion, the results from nearly all the metrics indicate a higher presence of careless and insufficient effort respondents in non-probability based panels. When we combine the results from each of the metrics, it becomes evident that, on average, non-probability based survey panels exhibit a higher percentage of careless and insufficient effort respondents compared to probability based survey panels. This provides strong evidence in support of the hypothesis that non-probability based panels have worse data quality when compared to probability based panels.

Reference

Carina Cornesse1,2 and Annelies G. Blom1,3. 2020. "Response Quality in Nonprobability and Probability-based Online Panels". Sociological Methods & Research. 52: 1-30. https://doi.org/10.1177/0049124120914940

Paul G. Curran. 2016. "Methods for the detection of carelessly invalid responses in survey data". Journal of Experimental Social Psychology. 66: 4-19.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686 https://doi.org/10.21105/joss.01686.

Xie Y (2023). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.35, https://github.com/rstudio/bookdown.

Wickham H, Miller E, Smith D (2023). *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*. R package version 2.5.3, https://CRAN.R-project.org/package=haven.

Zhu H (2021). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.3.4, https://CRAN.R-project.org/package=kableExtra.

Auguie B (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3, https://CRAN.R-project.org/package=gridExtra.

Tierney N, Cook D (2023). "Expanding Tidy Data Principles to Facilitate Missing Data Exploration, Visualization and Assessment of Imputations." *Journal of Statistical Software*, *105*(7), 1-31. doi:10.18637/jss.v105.i07 https://doi.org/10.18637/jss.v105.i07.

Xie Y (2023). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.43, https://yihui.org/knitr/.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Appendix: All code for this report

```
knitr::opts_chunk$set(
  echo = FALSE,
  eval = TRUE,
  message = FALSE,
  warning = FALSE,
  error = FALSE,
  out.width = "80%",
  fig.align = "center",
  fig.width = 8,
  fig.height = 7,
  fig.retina = 3,
  fig.pos = "H",
  out.extra = "",
  dev = "png")
# Load packages and data
library(tidyverse)
library(bookdown)
library(haven)
library(kableExtra)
library(gridExtra)
library(naniar)
library(knitr)

dat <- haven::read_sav("../Data and survey materials/ACSSM_Merged.sav")
# Visualize missing values in variables
dat %>%
  naniar::miss_var_summary() %>%
  head(10) %>%
  kable(caption = "Top 10 variables with a large number of missing values.") %>%
  kableExtra::kable_styling(latex_options = "hold_position")
# Count total variable with >10% missing values
miss10 <- dat %>%
  naniar::miss_var_summary() %>%
  filter(pct_miss > 10) %>%
  summarise("Count" = n()) %>%
  table() %>%
  kable(caption = "Total number of variables with more than 10% of missings") %>%
  kableExtra::kable_styling(latex_options = "hold_position")
# Summarize missing values in variables
miss_var_list <- dat %>% naniar::miss_var_summary()

miss_var_list <- miss_var_list[1:91,1]

dat <- dat[, !(names(dat) %in% miss_var_list$variable)]

# Mutate new column to indicate (non) probability based
dat1 <- dat %>%
  mutate(prob_based = ifelse(DataSource <= 5,
                             "Probablity based",
                             "Non-probablity based")) %>% # Indicate probability based
  filter(DataSource %in% c(2, 5, 6, 7, 8, 9)) # Filter data source
```

```r
# Mutate new column with DataSource name
dat1 <- dat1 %>%
  mutate(DataSource_name = case_when(
    DataSource == 2 ~ "Life in Australia",
    DataSource == 5 ~ "SMS push to web",
    DataSource == 6 ~ "Panel 1",
    DataSource == 7 ~ "Panel 2",
    DataSource == 8 ~ "Panel 3",
    DataSource == 9 ~ "Panel 4")) %>%
  mutate(DataSource_name = factor(DataSource_name,
                        levels = c("Life in Australia",
                                   "SMS push to web",
                                   "Panel 1",
                                   "Panel 2",
                                   "Panel 3",
                                   "Panel 4"),
                        labels = c("Life in Australia",
                                   "SMS push to web",
                                   "Panel 1",
                                   "Panel 2",
                                   "Panel 3",
                                   "Panel 4"))) # Use str() to check levels
# Plot time distribution
dat1 %>%
  mutate("Section_SUM_main" = Section_Loop_Section4_SectionTimeMinutes + # sum time of sections
           Section_Loop_Section5_SectionTimeMinutes +
           Section_Loop_Section6_SectionTimeMinutes +
           Section_Loop_Section7_SectionTimeMinutes +
           Section_Loop_Section9_SectionTimeMinutes) %>%
  ggplot(aes(x = DataSource_name,
                   y = Section_SUM_main,
                   color = prob_based)) +
  geom_point() +
  geom_boxplot() +
  scale_color_brewer(palette = "Set1") +
  theme_classic() +
  theme(axis.text.x = element_text(angle=15, hjust=1)) +
  labs(x = "",
       y = "Total time in minutes",
       color = "")
# Plot time distribution with 60 minutes cap
dat1 %>%
  mutate("Section_SUM_main" = Section_Loop_Section4_SectionTimeMinutes +
           Section_Loop_Section5_SectionTimeMinutes +
           Section_Loop_Section6_SectionTimeMinutes +
           Section_Loop_Section7_SectionTimeMinutes +
           Section_Loop_Section9_SectionTimeMinutes) %>%
  filter(Section_SUM_main < 60) %>%
  ggplot(aes(x = DataSource_name,
                   y = Section_SUM_main,
                   color = prob_based)) +
  geom_point() +
  geom_boxplot() +
```

```
  scale_color_brewer(palette = "Set1") +
  theme_classic() +
  theme(axis.text.x = element_text(angle=15, hjust=1)) +
  labs(x = "",
       y = "Total time in minutes",
       color = "",
       title = "Time taken < 60 minutes")


# Plot time distribution with 10 minutes cap
dat1 %>%
  mutate("Section_SUM_main" = Section_Loop_Section4_SectionTimeMinutes +
           Section_Loop_Section5_SectionTimeMinutes +
           Section_Loop_Section6_SectionTimeMinutes +
           Section_Loop_Section7_SectionTimeMinutes +
           Section_Loop_Section9_SectionTimeMinutes) %>%
  filter(Section_SUM_main < 10) %>%
  ggplot(aes(x = DataSource_name,
                  y = Section_SUM_main,
                  color = prob_based)) +
  geom_point() +
  geom_boxplot() +
  geom_hline(yintercept = 70 * 6 / 60, color = "purple", size = 1) +
  geom_text(aes(1.2,70 * 6 / 60, label = "`Cutoff` Time: 7 mins", vjust = 2),
            show.legend = FALSE,
            color = "purple") +
  scale_color_brewer(palette = "Set1") +
  theme_classic() +
  theme(axis.text.x = element_text(angle=15, hjust=1)) +
  labs(x = "",
       y = "Total time in minutes",
       color = "",
       title = "Time taken < 10 minutes")
knitr::include_graphics("Image/straight1.PNG")
# Straight lining, use variables: K6, TV_TIME, INTERNET, BENTYPE
# NOTE: 9 observations (`DataSource`: 2/6/6/6/7/7/7/8/9) select all `YES` in BENTYPE, only one obs refl
straight_lining <- dat1 %>%
  select(DataSource_name,
         prob_based,
         starts_with("K6_"),
         starts_with("TV_TIME_"),
         starts_with("INTERNET"),
         starts_with("BENTYPE_")) %>%
  rowwise() %>%
  mutate(K6_IER = ifelse(length(unique(c(K6_a_resp,
                                         K6_b_resp,
                                         K6_c_resp,
                                         K6_d_resp,
                                         K6_e_resp,
                                         K6_f_resp))) == 1, 1, 0)) %>%
  mutate(TV_TIME_IER = ifelse(length(unique(c(TV_TIME_a_resp,
                                              TV_TIME_b_resp,
                                              TV_TIME_c_resp,
```

```
                                            TV_TIME_d_resp,
                                            TV_TIME_e_resp,
                                            TV_TIME_f_resp))) == 1, 1, 0)) %>%
  mutate(INTERNET_IER = ifelse(length(unique(c(INTERNET_a_resp,
                                            INTERNET_b_resp,
                                            INTERNET_c_resp,
                                            INTERNET_d_resp))) == 1, 1, 0)) %>%
  mutate(BENTYPE_IER = ifelse(sum((c(BENTYPE_a_resp, # All `YES` for 5 BENTYPE => IER
                                     BENTYPE_b_resp,
                                     BENTYPE_c_resp,
                                     BENTYPE_d_resp,
                                     BENTYPE_e_resp
                                     ))) == 5, 1, 0)) %>%
  mutate(straight_lining = ifelse(sum(K6_IER,
                                  TV_TIME_IER,
                                  INTERNET_IER,
                                  BENTYPE_IER) >= 2, 1, 0))

straight_lining %>%
  select(DataSource_name, straight_lining) %>%
  table() %>%
  kable(caption = "'1' = Total number of 'Straight lining' respondents") %>%
  kable_classic() %>%
  kableExtra::kable_styling(latex_options = "hold_position")

# Plot a histogram of percentage of straight lining
straight_lining %>%
  select(prob_based, DataSource_name, straight_lining) %>%
  group_by(prob_based, DataSource_name) %>%
  summarise(Percentage = mean(straight_lining) * 100) %>%
  ggplot(aes(x = DataSource_name,
             y = Percentage,
             fill = prob_based)) +
  geom_col() +
  scale_fill_brewer(palette = "Set1") +
  theme_classic() +
  theme(axis.text.x = element_text(angle=15, hjust=1)) +
  labs(fill = "",
       x = "",
       title = "Percentage of straight lining in each panal") +
  geom_text(size = 6, aes(label = round(Percentage, 2)),
            vjust = -0.4,
            check_overlap = TRUE,
            position = position_dodge(width = 0.9)) +
  theme(legend.position = "right")
# Consistency: BENTYPE_d (Carer Allowance) == `YES` & UNPAIDCARE == `NO` => Not read the question caref
BENTYPE_UNPAIDCARE <- dat1 %>%
  select(DataSource_name,
         prob_based,
         BENTYPE_d_resp,
         UNPAIDCARE) %>%
  rowwise() %>%
  mutate(BENTYPE_UNPAID_IER = ifelse((BENTYPE_d_resp == 1) & (UNPAIDCARE == 2), 1, 0))
```

```r
# Plot percentages of inconsistent respondings
BENTYPE_UNPAIDCARE %>%
  group_by(prob_based, DataSource_name) %>%
  summarise(Percentage = sum(BENTYPE_UNPAID_IER) / n() * 100) %>%
  ggplot(aes(x = DataSource_name,
             y = Percentage,
             fill = prob_based)) +
  geom_col() +
  scale_fill_brewer(palette = "Set1") +
  theme_classic() +
  theme(axis.text.x = element_text(angle=15, hjust=1)) +
  labs(fill = "") +
  geom_text(size = 6, aes(label = round(Percentage, 2)),
            vjust = -0.5,
            check_overlap = TRUE,
            position = position_dodge(width = 0.9))

# Correct typo in variable name "FIRSTDRINK"
dat1 <- dat1 %>%
  mutate("FIRSTDRINK" = FIRSTDRINT)

# Plot a scatter-plot for variables 'AGE' AND 'FIRSTDRINK'
dat1 %>%
  select(DataSource_name,
         prob_based,
         AGE,
         FIRSTDRINK) %>%
  filter(AGE >= 0,
         FIRSTDRINK >= 0) %>%
  ggplot(aes(x = AGE,
             y= FIRSTDRINK)) +
  geom_point(col = "blue") +
  geom_point(data = dat1 %>% select(DataSource_name,
                                    AGE,
                                    FIRSTDRINK) %>%
               filter(AGE >= 0,
                      FIRSTDRINK >= 0,
                      FIRSTDRINK > AGE), color = "red", size = 3) +
  geom_abline(slope = 1, color = "red") +
  theme_classic() +
  geom_text(data = dat1 %>% select(DataSource_name,
                                    prob_based,
                                    AGE,
                                    FIRSTDRINK) %>%
               filter(AGE >= 0,
                      FIRSTDRINK >= 0,
                      FIRSTDRINK > AGE),
            size = 4,
            color = "red",
            aes(label = paste0(DataSource_name, "\n", "(", prob_based, ")")),
            vjust = -0.5,
            hjust = 0,
            check_overlap = TRUE,
```

```
            position = position_dodge(width = 0.9)) +
  xlim(c(0,125)) +
  ylim(c(0, 80)) +
  labs(y = "FIRSTDRINK",
       title = "Scatter plot of AGE and FIRSTDRINK")
# Plot time distribution for 'TV_TIME' variables
dat1 %>%
  select(DataSource_name,
         starts_with("TV_TIME")) %>%
  pivot_longer(TV_TIME_a_resp:TV_TIME_f_resp,
               names_to = "TV_TIME",
               values_to = "HOUR") %>%
  mutate(HOUR = ifelse(HOUR < 0, 0, HOUR)) %>%
  pivot_wider(names_from = TV_TIME,
              values_from = HOUR) %>%
  unnest() %>%
  mutate("TOTAL" = round((TV_TIME_a_resp +
                            TV_TIME_b_resp+TV_TIME_c_resp +
                            TV_TIME_d_resp+ TV_TIME_e_resp +
                            TV_TIME_f_resp) / 6 * 20), 0) %>%
  ggplot(aes(x = TOTAL)) +
  geom_histogram(fill = "#58a6f5")+
  theme_classic() +
  labs(title = "Minimum time spent on social media and TV services each week (in hours).",
       y = "Number of respondents",
       x = "Hour")
# Plot a histogram for long hour on video service and TV
dat1 %>%
  select(DataSource_name,
         prob_based,
         starts_with("TV_TIME")) %>%
  pivot_longer(TV_TIME_a_resp:TV_TIME_f_resp,
               names_to = "TV_TIME",
               values_to = "HOUR") %>%
  mutate(HOUR = ifelse(HOUR < 0, 0, HOUR)) %>%
  pivot_wider(names_from = TV_TIME,
              values_from = HOUR) %>%
  unnest() %>%
  mutate("TOTAL" = round((TV_TIME_a_resp +
                            TV_TIME_b_resp +
                            TV_TIME_c_resp +
                            TV_TIME_d_resp +
                            TV_TIME_e_resp +
                            TV_TIME_f_resp) / 6 * 20), 0) %>%
  mutate(TOTAL = ifelse(TOTAL > 112, 1 ,0)) %>%
  group_by(prob_based, DataSource_name) %>%
  summarise(Percentage = sum(TOTAL) / n() * 100) %>%
  ggplot(aes(x = DataSource_name,
             y = Percentage,
             fill = prob_based)) +
  geom_col() +
  scale_fill_brewer(palette = "Set1") +
  theme_classic() +
```

```
  labs(fill = "") +
  geom_text(size = 4,
            aes(label = round(Percentage, 2)),
            vjust = -0.5,
            check_overlap = TRUE,
            position = position_dodge(width = 0.9)) +
  theme(axis.text.x = element_text(angle=15, hjust=1))
# Plot percentage of selection middle answer in each survey for a selection of questions
dat1 %>%
  select(DataSource_name,
         prob_based,
         starts_with("K6_")) %>%
  pivot_longer(K6_a_resp:K6_f_resp,
               names_to = "K6_name",
               values_to = "K6_value") %>%
  mutate(K6_value = as.numeric(K6_value)) %>%
  mutate(K6_IER = ifelse(K6_value == 3, 1, 0)) %>%
  group_by(prob_based, DataSource_name) %>%
  summarise(Percentage = sum(K6_IER) / n() * 100) %>%
  ggplot(aes(x = DataSource_name,
             y = Percentage,
             fill = prob_based)) +
  geom_col() +
  scale_fill_brewer(palette = "Set1") +
  theme_classic() +
  labs(fill = "",
       x = "",
       title = "Percentage of Mid-point selection for selected questions") +
  geom_text(size = 5,
            aes(label = round(Percentage, 0)),
            vjust = -0.5, check_overlap = TRUE,
            position = position_dodge(width = 0.9)) +
  theme(legend.position = "bottom")
# Percentage of item non-response
Item_non_response <- dat1 %>%
  select(c(DataSource_name,
           prob_based,
           GENDER:SUBURB,
           IMPPROB:INCOME,
           -GENDER_4,
           -SUBURB,
           -IMPPROB,
           -VOTE_PARTY_96,
           -HOMEOWNER_7,
           -PANELREASON_96,
           -HIGHEST_QUALIFICATION_7,
           -COB_7)) %>%
  as.data.frame() %>%
  pivot_longer(GENDER:INCOME,
               names_to = "Question",
               values_to = "Answer") %>%
  mutate(Answer = as.numeric(Answer)) %>%
  mutate(Non_response = ifelse(Answer %in% c(-98, -99), 1, 0)) %>%
```

```
  select(DataSource_name, prob_based, Non_response) %>%
  group_by(prob_based, DataSource_name) %>%
  summarise(Percentage = sum(Non_response) / n() * 100)

Item_non_response %>%
  ggplot(aes(x = DataSource_name,
             y = Percentage,
             fill = prob_based)) +
  geom_col() +
  scale_fill_brewer(palette = "Set1") +
  theme_classic() +
  labs(fill = "",
       x = "",
       title = "Percentage of selecting 'Don't know' or 'Refused to answer'") +
  theme(axis.text.x = element_text(angle=15, hjust=1)) +
  geom_text(size = 5,
            aes(label = round(Percentage, 2)),
            vjust = -0.5,
            check_overlap = TRUE,
            position = position_dodge(width = 0.9)) +
  ylim(c(0,1))


# Combine all metrics
# Identify speeding respondents
speed <- dat1 %>% select(DataSource_name,
                Section_Loop_Section4_SectionTimeMinutes:Section_Loop_Section9_SectionTimeMinutes,
                -Section_Loop_Section8_SectionTimeMinutes,
                prob_based) %>%
  rowwise() %>%
  mutate("ave" = sum(c(Section_Loop_Section4_SectionTimeMinutes +
           Section_Loop_Section5_SectionTimeMinutes +
           Section_Loop_Section6_SectionTimeMinutes +
           Section_Loop_Section7_SectionTimeMinutes +
           Section_Loop_Section9_SectionTimeMinutes), na.rm = TRUE) *60 /70) %>%
  mutate(ave = ifelse(ave < 7, 1, 0)) %>%
  select(ave)

# Identify straight lining respondents
straight <- dat1 %>%
  select(DataSource_name,
         starts_with("K6_"),
         starts_with("TV_TIME_"),
         starts_with("INTERNET"),
         starts_with("BENTYPE_")) %>%
  rowwise() %>%
  mutate(K6_IER = ifelse(length(unique(c(K6_a_resp,
                                         K6_b_resp,
                                         K6_c_resp,
                                         K6_d_resp,
                                         K6_e_resp,
                                         K6_f_resp))) == 1, 1, 0)) %>%
  mutate(TV_TIME_IER = ifelse(length(unique(c(TV_TIME_a_resp,
```

```
                                          TV_TIME_b_resp,
                                          TV_TIME_c_resp,
                                          TV_TIME_d_resp,
                                          TV_TIME_e_resp,
                                          TV_TIME_f_resp))) == 1, 1, 0)) %>%
  mutate(INTERNET_IER = ifelse(length(unique(c(INTERNET_a_resp,
                                          INTERNET_b_resp,
                                          INTERNET_c_resp,
                                          INTERNET_d_resp))) == 1, 1, 0)) %>%
  mutate(BENTYPE_IER = ifelse(sum((c(BENTYPE_a_resp, # All `YES` for 5 BENTYPE => IER
                                     BENTYPE_b_resp,
                                     BENTYPE_c_resp,
                                     BENTYPE_d_resp,
                                     BENTYPE_e_resp
                                     ))) == 5, 1, 0)) %>%
  mutate(straight_lining = ifelse(sum(K6_IER,
                                  TV_TIME_IER,
                                  INTERNET_IER,
                                  BENTYPE_IER) >= 2, 1, 0))

# Identify inconsistent answer respondents
consistency1 <- dat1 %>%
  select(DataSource_name,
         prob_based,
         starts_with("BENTYPE")) %>%
  rowwise() %>%
  mutate(K6_IER = ifelse(sum((c(BENTYPE_a_resp,
                                BENTYPE_b_resp,
                                BENTYPE_c_resp,
                                BENTYPE_d_resp,
                                BENTYPE_e_resp
  ))) == 5, 1, 0))


consistency2 <- dat1 %>%
  select(DataSource_name,
         prob_based,
         BENTYPE_d_resp,
         UNPAIDCARE) %>%
  rowwise() %>%
  mutate(BENTYPE_UNPAID_IER = ifelse((BENTYPE_d_resp == 1) & (UNPAIDCARE == 2), 1, 0))


consistency3 <- dat1 %>%
  select(DataSource_name,
         prob_based,
         AGE,
         FIRSTDRINK) %>%
  mutate(AGE = ifelse(AGE < 0, 0, AGE), # set 0
         FIRSTDRINK = ifelse(FIRSTDRINK < 0, 0, FIRSTDRINK)) %>%
  mutate(AGE_FIRSTDRINK = ifelse(AGE < FIRSTDRINK, 1, 0))

# Identify outlier from TV_TIME
```

```r
extreme <- dat1 %>%
  select(DataSource_name,
         starts_with("TV_TIME")) %>%
  pivot_longer(TV_TIME_a_resp:TV_TIME_f_resp,
               names_to = "TV_TIME",
               values_to = "HOUR") %>%
  mutate(HOUR = ifelse(HOUR < 0, 0, HOUR)) %>%
  pivot_wider(names_from = TV_TIME,
              values_from = HOUR) %>%
  unnest() %>%
  mutate("TOTAL" = round((TV_TIME_a_resp +
                            TV_TIME_b_resp +
                            TV_TIME_c_resp +
                            TV_TIME_d_resp +
                            TV_TIME_e_resp +
                            TV_TIME_f_resp) / 6 * 20, 0)) %>%
  mutate(TT = ifelse(TOTAL > 112, 1, 0))

# Combine results
dat1 %>%
  mutate(flag_speed = speed$ave,
         flag_healthcon98 = as.numeric(HEALTHCON13),
         flag_healthcon99 = as.numeric(HEALTHCON14),
         flag_straight = straight$straight_lining,
         flag_consistency1 = consistency1$K6_IER,
         flag_consistency2 = consistency2$BENTYPE_UNPAID_IER,
         flag_consistency3 = consistency3$AGE_FIRSTDRINK,
         flag_extreme = extreme$TT) %>% # flag_consistency3,
  mutate(sum_flag = flag_speed + flag_healthcon98 + flag_healthcon99 +
           flag_straight + flag_consistency1 + flag_consistency2 + flag_extreme) %>%
  mutate(combine_flag = ifelse(sum_flag > 0, 1, 0)) %>%
  select(DataSource_name, prob_based, combine_flag) %>%
  group_by(prob_based, DataSource_name) %>%
  summarise(percentage = mean(combine_flag)) %>%
  group_by(prob_based) %>%
  summarise(Percentage = round(mean(percentage) * 100, 2))  %>%
  kable(caption = "Average percentage of careless and insufficient effort respondents") %>%
  kableExtra::kable_styling(latex_options = "hold_position")
```