NCAER | SINCE 1956 | QUALITY. RELEVANCE. IMPACT

NCAER National Data Innovation Centre

Institutional Research Grant Report Number 04, October 2021

# Patterns in Paradata from Delhi Metropolitan Area Study

Arpita Ghosh and Laurent Billot

# PATTERNS IN PARADATA FROM DELHI METROPOLITAN AREA STUDY

***Arpita Ghosh***
Senior Research Fellow, Biostatistics and Data Science, George Institute India
Conjoint Lecturer, University of New South Wales, Sydney, NSW, Australia
Professor, Prasanna School of Public Health,
Manipal Academy of Higher Education
aghosh@georgeinstitute.org.in

***Laurent Billot***
Director, Biostatistics and Data Science, The George Institute for Global Health Australia
Associate Professor, Faculty of Medicine, University of New South Wales, NSW, Australia
lbillot@georgeinstitute.org

**Acknowledgement:** The authors would like to thank Sonalde Desai, Santanu Pramanik, Bijay Chouhan, and Arpita Kayal at the NCAER National Data Innovation Centre (NDIC) for help with survey data and paradata, for providing suggestions and for operational support. They wish to thank the reviewers for careful review of the midterm report and detailed feedback. This research is supported by the National Council of Applied Economic Research through the NCAER NDIC. The authors are thankful to NCAER NDIC for the opportunity to work on this novel data.

**Funding Support and Disclaimer:** This research is supported by the National Council of Applied Economic Research through the NCAER National Data Innovation Centre. The views presented in this paper are those of the authors and not those of NCAER or its Governing Body. Funding for the NCAER National Data Innovation Centre is provided by Bill & Melinda Gates Foundation.

**Suggested Citation:**
Ghosh, Arpita and Laurent Billot. (2021). "Patterns in Paradata from Delhi Metropolitan Area Study", *Institutional Research Grant Report Number 04*. Report submitted to NCAER National Data Innovation Centre, New Delhi: National Council of Applied Economic Research.

**TABLE OF CONTENTS**

## LIST OF FIGURES AND TABLES

# 1. Introduction

Couper in 1998 first introduced the term 'paradata' to refer to survey process data in the field of survey methodology (Groves and Couper 2012). In surveys using computer-assisted personal interviewing (CAPI) software programmes, such as Blaise, a huge amount of process data are generated throughout the survey. This may include interviewer productivity indicators, call records, number of attempts made to interview the targeted respondent, interview length, item-level time stamp data based on key strokes, use of item-specific remarks, GPS coordinates, and audio recording of interviews. The scope of paradata can be expanded to also include observational information, for example, observed neighbourhood conditions, observations made by interviewers about respondents, and so on.

Systematic and timely examination of the paradata can shed light on sources of error, such as non-response error and measurement error and help improve survey data quality (Kreuter 2013). Paradata can also be used to reduce non-response in a responsive design framework, that is, use paradata to monitor data collection in real time and provide interventions for subsequent waves of data collection. Methodological research using paradata has mainly focused on non-response error. Research on the comprehensive use of measurement-error-related paradata in surveys, accounting for the complex, hierarchical data structure, is limited.

We attempt to fill this gap by examining how paradata can be used to detect differences in the interview process. Studies investigating interview quality using the entire range of available paradata are limited. We use cluster analysis, a novel application in this context, to uncover any underlying patterns in paradata. To achieve our objective we use paradata from the Delhi Metropolitan Area Study (DMAS) baseline survey, conducted by the National Council of Applied Economic Research. We identified patterns in DMAS baseline paradata and examined how the patterns are associated with interview characteristics. We hope that the learnings from this exercise will feed into other surveys being conducted using computer-assisted methods.

# 2. Data

## 2.1. Survey Data

DMAS was carried out during the period February 15–June 3, 2019 in the National Capital Region (NCR) of India comprising of 31 districts spread across four States—the National Capital Territory (NCT) of Delhi (9 districts), Rajasthan (2 districts), Uttar Pradesh (7 districts), and Haryana (13 districts). The survey gathers information on participants on different domains including household income and expenditure, labour force participation, financial inclusion, health insurance and healthcare expenditure, gender equality and

empowerment, among others. A three-stage stratified cluster sampling design was used to ensure random sampling at each stage of sample selection – districts in the first stage, villages in rural areas and NSSO UFS blocks in urban areas in the second stage, and households in the third stage of sample selection. A total of 27,456 individuals in 5255 households in both rural and urban areas were included in the survey. Data were collected through CAPI using Blaise software.

## 2.2. Paradata

The Blaise audit trail data records an interviewer's interaction with the questionnaire. It shows us what the interviewer did on the field, which questions the interviewer asked, and how long each event took. Many special actions are also recorded, such as edits, making remarks, etc. We use 'item' to indicate each survey field that was captured in the Blaise data model. The item-level paradata indicators that were recorded in the DMAS survey and were analysed in this exercise included the:

- Number of times that the variable (question) was visited by the interviewer during the interview;
- Distinct separate answers recorded for the variable during the interview;
- Longest visit duration for the variable during the interview;
- Combined visit duration time of all visits to the variable during the interview;
- Number of times the interviewer used the 'Remark' feature for the variable;
- Number of times answering the question (variable) raised a soft-check in the Blaise data model;
- Number of times an interviewer selected 'Don't Know' for the variable during the interview;
- Number of times the interviewer selected 'Refused' for the variable;
- Number of times the interviewer selected 'Quit' for the variable;
- Number of visits to the variable during the interview with a field duration longer than 5 minutes; and
- Combined duration of all visits to variable that lasted longer than 5 minutes.

## 2.3. Non-paradata Indicators

In addition to paradata, DMAS also collected data on interviewer characteristics. We considered interviewer background data and a few characteristics of interviews or households that may help explain the observed patterns in paradata. The list of interviewer and household characteristics that were examined is presented in the following table:

| Interviewer Characteristics | Interview Characteristics |
|---|---|
| 1. Sex<br>2. Age<br>3. Education<br>4. Religion<br>5. Caste<br>6. Type of area (rural/urban) in which interviewer grew up<br>7. Whether interviewer/family involved in agriculture<br>8. Knowledge of word, excel, email and internet<br>9. Whether was involved in IHDS survey | 1. State<br>2. Type of area (rural/urban)<br>3. Household size<br>4. Household wealth quintile<br>5. Religion and caste of head of household<br>6. Time (week/month of survey period) |

# 3. Results

The audit trail file contains paradata on interviews conducted in 5231 households by 29 interviewers. The households were asked a variable number of questions, depending on several factors, including the number of household members, number of women in the reproductive age group, number of children, the primary economic activity of the household members, and so on. Therefore, item-level paradata are available for a varying number of items per interview, ranging between 298 and 1105 items. For this analysis, we focus on a sub-set of 203 items for which paradata were available for all 5231 interviews. These 203 items are spread across the different modules in the questionnaire. Of the 203 items, we dropped three items, that is, the display status table, consent for interview, and consent for recording, and considered items belonging to Sections 1 to 30.

## 3.1. Item-wise Summary

We obtained summaries for interviews for different paradata indicators, for the 200 items. For each item, we calculated the number of interviews where the item was visited multiple times during the interview (yes/no), the response was revised (yes/no), and the number of visits (over all interviews) where the options 'Remark' 'Quit', 'Don't' Know' or 'Refused' were selected, and a soft check was triggered. We derived the median and interquartile range of item-level duration (combined visit duration over all visits to variable) over all interviews. We also calculated, for each item, the total number and duration of pauses over all interviews, and the number of interviews where maximum duration was less than 2 seconds.

These summaries are presented in Figure 1. The items are arranged by different sections in the questionnaire and are indicated using different colours. We note that the consumption expenditure/assets section stands out—the items in this section were visited

multiple times, the responses were revised, the median duration was higher, and the options 'Remark' and 'Don't Know' were more frequently selected by interviewers.

## 3.2. Interview-wise Summary

We next derived interview-level summaries of the item-level paradata indicators. We used the 200 items that were common across all 5231 interviews. The interview-level summaries included the:

- Proportion of items out of 200 that were visited more than once during the interview;
- Proportion of items out of 200 for which the response was revised one or more times during the interview;
- Duration (in seconds) per item, i.e., median (over 200 items) of duration (combined visit duration of all visits to variable during the interview);
- Proportion of items that take more time than average, i.e., proportion of these 200 items with duration more than median duration (over 5231 interviews) for that item;
- Proportion of items with the longest visit duration for the item during interview less than 2 seconds; and
- Number of times the option 'Remark', 'Quit', 'Don't' Know' or 'Refused' was selected, and soft check was triggered during the interview (over all visits to 200 items)

We plot below the distribution of these variables in Figure 2. We plot densities for the first five variables and histograms of the count data for the last five.

We note that the median percentages of items (out of 200) that were visited more than once, and for which the response was revised during the interview, are 8% and 4.5%, respectively. However, in 10% of the interviews, 21% of the items were visited more than once during the interview and 9.5% of the items with responses revised once or more times during the interview. The median per-item duration was 3.5 seconds and 10% of the interviews had a per-item duration of 5.3 seconds or more. While 15% of the interviews had 28.5% or more items with the longest visit duration of less than 2 seconds, 24%, 19%, and 2% of the interviews selected the options, 'Remark', 'Don't Know' and 'Refused' at least once during interview. Further, 5.6% of the interviews had one or more pauses during the interview and their median total pause duration was 7.2 minutes.

## 3.3. Cluster Analysis

We performed cluster analysis to identify sub-groups of interviews based on paradata. Clustering helps in organising things (interviews in our case) that are *close* into groups (Chavent, Kuentz et al. 2011). For this, we need to define distance/similarity between two interviews using the characteristics (paradata indicators in our case) of interviews.

We used all available paradata for 200 items to identify clusters among 5231 interviews. We had to drop 11 interviews due to missing data. We used the following 11 paradata indicators: number of times the item was visited during the interview, number of

responses recorded, total duration over all visits, whether the maximum visit duration was less than 2 seconds, number of pauses, total pause duration, number of times a soft check was triggered, and number of times the options, 'Don't Know', 'Refused', 'Remark', and 'Quit' were selected. So, we clustered 5220 interviews on the basis of 2200 variables or features.

To decide on the distance/similarity measure between two interviews, we need to consider the variable types. The variables were of mixed type—*duration* was continuous, *whether longest visit duration was less than 2 seconds* was binary, and the rest were count data. There are few options available for distance measure when there are mixed-type variables. The most popular distance for mixed-type variables is derived as the complement of the Gower's similarity coefficient—an average of the distances calculated variable by variable, the single distances are all scaled to range from 0 (minimum distance) to 1 (maximum distance).

We used Gower distance as the metric to calculate distances between interviews based on the 2200 variables. Before calculating the distance, we took logarithmic transformation of the *duration* variables. We then applied Partitioning Around Medoids(PAM) algorithm to cluster the interviews (Budiaji and Leisch 2019). This is similar to the popular k-means clustering algorithm, except that "means" as the centre of the clusters (centroid) are replaced with medoids as the robust representation of the cluster centres. This is important in the common context that many points do not belong well to any cluster.

PAM clusters the data set of n objects into k clusters known *a priori*. We have to specify the number of clusters and the algorithm identifies the best partition of the data into the specified number of clusters. We used the silhouette criteria to decide the number of clusters (Akhanli and Hennig 2020). The silhouette coefficient contrasts with the average distance to elements in the same cluster with the average distance to elements in other clusters. Objects with a high silhouette value are considered as well-clustered, while objects with a low value may be outliers. This index works well with PAM, and is used to determine the optimal number of clusters. The middle plot in Figure 3 presents the silhouette coefficients for the number of clusters from 2 to 8. We note that the optimal number of clusters is 3 (the objective is to maximise silhouette width and choose a parsimonious model).

We use the PAM algorithm to identify the 3 clusters of sizes 3763, 759, and 698 interviews. We then visualise the identified clusters. We used t-SNE (t-distributed Stochastic Neighbor Embedding) to visualise the clusters in a two-dimensional space. t-SNE is a non-linear dimensionality reduction technique, that is, this algorithm allows us to separate data that cannot be separated by any straight line. The first plot in Figure 3 simply presents a two-dimensional mapping of the multi-dimensional (2200 dimensions in our case) feature space. The rightmost plot in Figure 3 is the two-dimensional mapping with three colours representing the identified and well-separated clusters—grey represents the largest cluster with 3763 interviews, and the red and blue dots represent the two smaller (759 and 698 interviews, respectively) but well-separated clusters.

## 3.4. Characterisation of Clusters

### *3.4.1. Paradata Characteristics*

We next wanted to explore how the features/variables that went into the construction of the clusters are distributed across the three clusters, i.e., define the clusters with respect to paradata. From Figures 1 and 2, we saw that the item-level duration varies considerably across interviews and also across items, and may therefore, have helped partition the interviews during cluster analysis. Figure 4 shows the average standardised log duration for each item, by clusters. We see clear patterns with respect to item-level duration—Cluster 2 is comprised of interviews that spent more time on all items, while Cluster 3 is comprised of interviews that spent less time on items, in general. This characterisation depends on the number of clusters. We note from the middle panel of Figure 3 that we can select up to six clusters without a substantial drop in silhouette width. We therefore, as part of sensitivity analyses, present the above figure (average standardised log duration for each item, by clusters) for six clusters in the Appendix (Supplementary Figure 4). We, however, run the risk of over-fitting with six clusters. The decision as to which of these clusters make more sense or are useful for understanding interview quality is a combination of both domain knowledge and statistical criteria. Based on these combined considerations, we continue with three clusters characterised as above.

We now examine the distribution of other paradata indicators across the three clusters (Figure 5). For this, we have summarised the paradata indicators over items. We observe that the median per-item time is 5.4 seconds for interviews in Cluster 2 as compared to 2.6 seconds for interviews in Cluster 3. This translates to a difference of 24 minutes for an interview of average length (median number of items is 513). Also, the median percentage of 200 items that were covered under 2 seconds (the maximum visit time was less than 2 seconds) was 3% in Cluster 2 as compared to 34% in Cluster 3. In Cluster 3 (n=698), 18 more interviews selected the option 'Refused' and 74 more interviews selected the option 'Don't Know' as compared to Cluster 2 (n=759). Plots using paradata for all items in the interview (Appendix Figure S2) show a similar pattern.

### *3.4.2. Interview and Interviewer Characteristics*

We also want to characterise the three clusters with respect to interview/household characteristics and interviewer characteristics. We first examine how the clusters are distributed across the interviewers and the date of interview (Figure 6). The first plot in the panel is a scatter plot of interviews by interviewer (y-axis) and date of survey (x-axis). The colours present the cluster membership—the largest cluster is in grey, and the two smaller clusters are in red and blue. We saw from Figures 4 and 5 that the red and blue clusters are comprised of interviews that generally speaking, take more time and less time, respectively. The interviewers on the y-axis are arranged in order such that the interviewers at the top have the highest proportion of Cluster 2 interviews.

We observe strong clustering of cluster membership by both interviewers and the week/month of the survey period. Clearly, the interviews in the beginning take longer and belong more often to Cluster 2 (denoted by red). Over time, the interviews get shorter, that is, signifying the transition from red to grey to blue. However, this pattern varies from interviewer to interviewer. Some interviewers transition more slowly than others and the longer-duration interviews continue even after the first month for some interviewers.

We wanted to compare if instead of cluster membership we simply categorise the interview duration, whether we would observe similar patterns. In the top right plot, we have the exact same scatter plot, except that the colours now are based on categories of total time for the 200 common items. Three categories were formed such that they contain the same number of interviews as the three clusters, that is, 698 interviews that took the least time: 8-17 minutes (denoted by blue), 3763 interviews that took 17-35 minutes (denoted by grey), and 759 interviews that took the most time: 35-81 minutes (denoted by red) to complete the 200 items. The general pattern is similar—longer interviews in the beginning and the interviews getting shorter over time. But many interviews are classified differently according to this criteria. Therefore, we next compared the total time for 200 items with cluster membership (bottom row plot in Figure 6) by survey date. The interviews in the beginning of the survey clearly take much longer and belong to Cluster 2. But as time goes on, the distinction between interviews simply based on total time (summary of item-level duration—a single dimension) does not entirely match with cluster membership (based on mutiple dimensions—item-level duration and other paradata).

While it is clear that duration decreases over the months, it is difficult to identify whether this is because the survey started out with rural areas, branching out to urban areas later, or whether the interview time reduced with experience over the survey period. To have some idea, we plot interview duration over months by type of residence (bottom row Supplementary Figure 3). We also present a graph exploring the interview duration by the time of the day. The hypothesis was that there may be rush interviewing during a particular time of the day, especially towards the end. We, therefore, examined the pattern in interview duration (distribution of interviews across the three clusters) by date and time (top row Supplementary Figure 3). We later model item the response time as a function of these factors to further explore these associations.

We next wanted to characterise the clusters with respect to household and interviewer characteristics. Table 1 presents the distribution of the interviewed households and the interviewers that conducted these interviews. We present in Figure 7 the distribution of different household characteristics of interviews in the three clusters. This may help explain the differences in paradata that we observe across the three clusters. We examine the distribution of household size, state, type of residence, household wealth quintile, religion, and caste of the household head across the interviews in the three clusters. Household size is similarly distributed with median size 5. Cluster 2 had more interviews conducted in Uttar Pradesh and in rural areas—this makes sense as the survey started in rural areas of Uttar Pradesh. This also explains why Cluster 2 has a higher proportion of poorest households. Cluster 2 had the highest proportion of households with heads belonging to the Other Backward Caste category. Cluster 3 had more interviews

conducted in Delhi and Haryana, and in urban areas. It had the lowest proportion of poorest households and households with heads belonging to Other Backward Class, the highest proportion of richest households, and the highest proportion of households with heads belonging to the General/Forward caste.

Figure 8 describes how the three clusters are different with respect to interviewer characteristics. The interviews in Cluster 2 were conducted by interviewers who are more likely to be older, belonging to the Other Backward Caste or Scheduled Caste categories, be a Hindu and have a Master's degree. They are also more likely to be male and either do or have a family that does farming. In comparison, the interviews in cluster 3 were conducted by interviewers who are likely to be younger, female, and have grown up in a rural area. Among them the highest share of interviewers was of those belonging to the Sikh community and to the Scheduled Caste, and the lowest percentage of interviewers were those with a Master's degree. They also had the lowest percentage reporting knowledge of Excel, but has the highest percentage reporting proficiency in Internet use.

## 3.5. Variation in Item-level Response Times

We analysed item-level response times (time taken to answer a question) as a function of item level characteristics, household/interview level characteristics, and interviewer level characteristics (Elliott and West 2015). Using multilevel models, we explore how these different factors influence item-level response times (Couper and Kreuter 2013). In this exercise, we use all items instead of the 200 items we restricted our attention to earlier. The item-level characteristics were automatically derived question characteristics from Blaise audit trails and the data model. The household/interview and interviewer characteristics have been discussed earlier. These are presented in Table 3.

We fit a series of multilevel linear mixed models starting with the null model including random effects for interview/household, interviewer, week, and neighborhood (village/town). The null model can be specified as

$$y_{ij} = \beta_0 + u_j^{(2)} + u_{iwer(j)}^{(3)} + u_{wk(j)}^{(4)} + u_{neigh(j)}^{(5)} + \epsilon_{ij}$$

where $y_{ij}$ is the logarithm of time taken for item *i* in interview *j* – conducted by interviewer *iwer(j)* in week *wk(j)* in household belonging to village/town *neigh(j)*. In the above model, $u_j^{(2)}$, $u_{iwer(j)}^{(3)}$, $u_{wk(j)}^{(4)}$ and $u_{neigh(j)}^{(5)}$ represent random effects associated with the interview/household, interviewer, week, and neighborhood (village/town), respectively, and $\epsilon_{ij}$ represents the residual variability that is associated with each item *i*. All random effects and residual errors are assumed to be normally distributed.

We then added covariates successively to this model to assess the sources of variation. We first included item-level characteristics, then added household/interview-level characteristics and finally interviewer-level characteristics. The full model can be specified as

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 v_j + \beta_3 z_{iwer(j)} + u_j^{(2)} + u_{iwer(j)}^{(3)} + u_{wk(j)}^{(4)} + u_{neigh(j)}^{(5)} + \epsilon_{ij}$$

where $x_{ij}$ are present item-level characteristics, $v_j$ present interview-level characteristics and $z_{iwer(j)}$ present interviewer-level characteristics.

The estimates of variance components and intra-class correlation coefficients (ICCs) from the different models are presented in Table 2. The magnitudes of variance components are then compared to assess the relative contribution of each level to the variation in item response time. From the ICCs from null model it can be seen that interviewers contribute around 2.3% of the total variation and households/interviews contribute 2.2% of the variation. Most of the variation (91%) is at the item level. Adding the item-level characteristics in Model 1 accounted for about 19.3% of the variation at the item level. However, adding interview/household-level characteristics in Model 2 and further adding interviewer-level characteristics in Model 3, led to modest reductions in variation—2.7% and 0.5%, respectively.

Table 3 presents the fixed effect estimates from the final multilevel model. Questions that are longer require more time. Questions that are open-ended take the longest time. Multiple-choice questions and those requiring a numeric response also require more time than single-choice questions. Similarly, questions having interviewer instructions take longer. Sequence number is negatively associated, suggesting that the time to administer a question decreases over the course of the interview. Household wealth quintile is associated with response times—the ones that are neither in the poorest nor in the richest quintiles take longer, most likely due to complexity in capturing their economic activity and assets. Similarly, item response time is higher in households engaged in farming activities and for households where longer questionnaires (including more complex questions that require probing and therefore more scheduled time) are administered. The item response time decreased over the months of the survey period as interviewers got more familiar with the questionnaire. The association with the time of the interview suggests that interviews conducted in the late afternoon are likely to be finished more quickly, possibly due to rushing on the part of the interviewers.

We noted earlier that interviewers accounted for a small proportion of the variation in response times. Having a Master's degree and belonging to the Scheduled Caste or Forward/General Caste categories were found to be associated with higher item response times. In general, adding interviewer-level characteristics did not help explain much of the interviewer-level variation.

## 3.6. Employment-based Quality Indicator

Finally, we assessed how the cluster membership was associated with quality indicators derived based on survey data. We compared the general question on employment with detailed responses from subsequent sections on the involvement of household members in farming activities, tending to livestock, business activities, etc. The concordance between the two sets of responses indicated a good-quality interview. This indicator is based on the observation that the household respondent generally underplays the employment status for women, while for men, the employment status section often

records a 'yes' whereas the subsequent detailed questions indicate otherwise. The individual-level responses were combined to create a binary variable at the household level, indicating whether there was a discrepancy in employment record for any member. In Figure 9, we present the proportion of 'poor quality' interviews across the three clusters by the month of the survey. The four plots correspond to how employment status was defined for the individual based on reported activities: 1) reported working for at least 1 day in either wage and salary, or farm or non-farm sections of activity listing, 2) reported working for at least 30 days in either wage and salary, or farm or non-farm sections of activity listing, 3) reported working for at least 30 days combined across wage and salary, farm and non-farm sections of activity listing and 4) reported working for at least 30 days combined across wage and salary, farm, non-farm and animals (livestock) sections of activity listing. The plots suggest that the proportion of 'poor-quality' interviews was lower in Cluster 3 than in Clusters 1 and 2 in the first one-and-a-half months of the survey. In the next two months, the proportion of 'poor-quality' interviews in Clusters 1 and 2 went down and was low across all the three clusters, perhaps because of the feedback provided to the interviewers.

## 4. Conclusion

In conclusion, it may be pointed out that we have demonstrated that paradata can shed light on how the interviews were carried out, which interviewers may be struggling, and which sections of the questionnaire may need more thought or training. We should further look into quality indicators and how the clusters are associated with different quality indicators. An alternative quality indicator can be defined based on medical expenditure in the last 30 days in the consumption expenditure section and similar questions (presumably more detailed ones) from the health expenditure section. A continuous quality indicator will have more information than a binary quality indicator and may help unpack the contribution of paradata in explaining the variation in quality. We can also calculate the negative screening rate for each interview and use that as a quality indicator. However, these are all attempts to define the quality of an interview automatically using survey data. The definition of a robust quality indicator is challenging without manual coding of the interviews to assess quality. Perhaps this can be attempted for a sub-sample of the interviews and the survey data-based quality indicators can then be assessed against this measure.

We have analysed item-level response times as a function of item-level characteristics, household/interview characteristics and interviewer characteristics. The hierarchical and cross-classified factors (items clustered within interviews and interviews clustered within the two cross-classified factors – interviewers and survey period) have been accounted for within a multilevel framework. We have demonstrated that item-level features automatically derived from audit trail files and data models can help explain the variation in response times. However, much of the variation in item response times remains unaccounted for. Household, interview, and interviewer features contributed modestly to the overall variation. The residuals from this analysis can shed light on items, interviews, and interviewers who take more or less time than expected and these 'outliers' can then be further investigated.

# REFERENCES


Akhanli, S.E. and C. Hennig (2020). "Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes." *Statistics and Computing*, 30(5): 1523-1544.

Budiaji, W. and F. Leisch (2019). "Simple K-medoids partitioning algorithm for mixed variable data." *Algorithms*, 12(9): 177.

Chavent, M., V. Kuentz, B. ı. Liquet and L. Saracco (2011). "ClustOfVar: An R package for the clustering of variables." *arXiv preprint arXiv:1112.0295*, 50(13): 1-16.

Couper, M.P. and F. Kreuter (2013). "Using paradata to explore item level response times in surveys." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(1): 271-286.

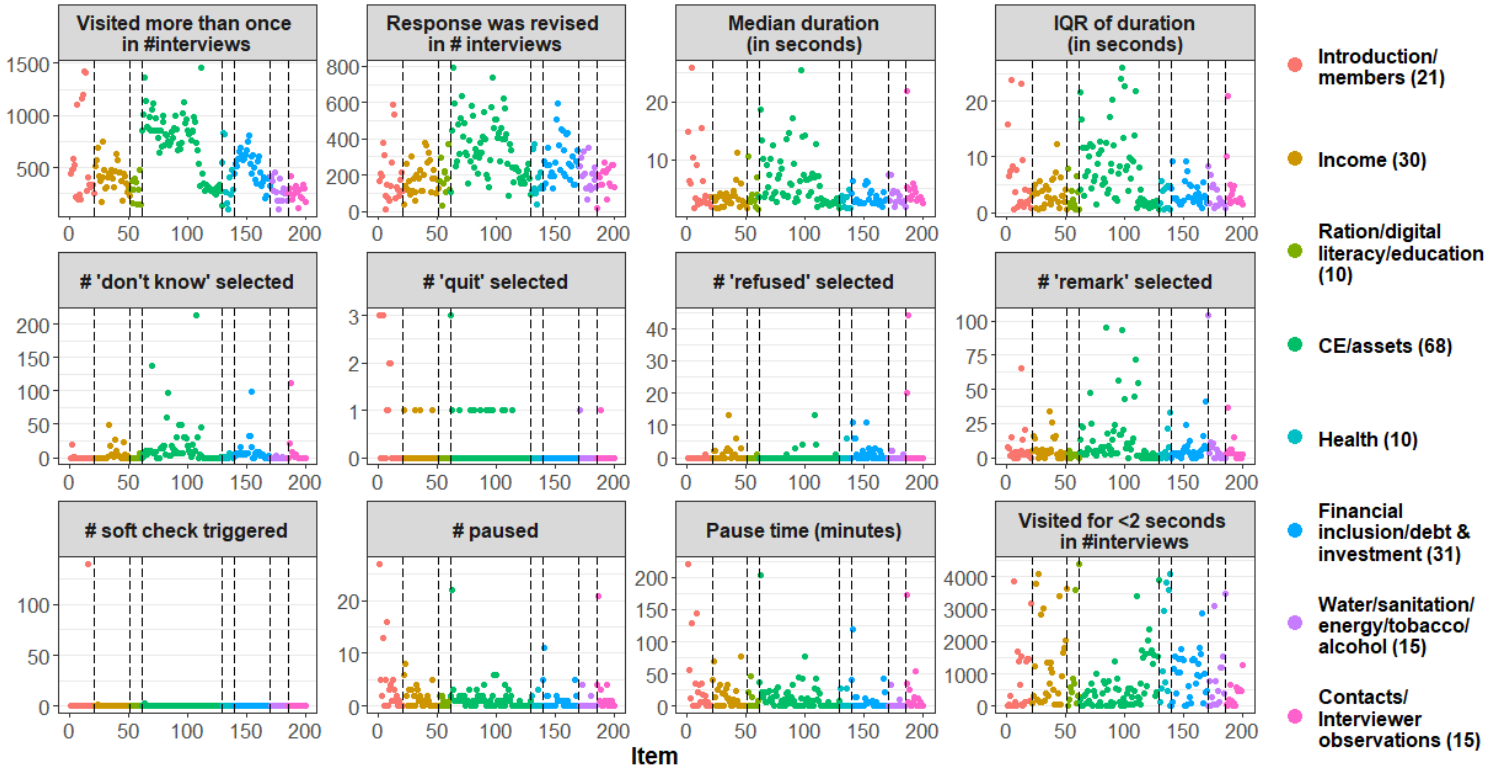
Elliott, M.R. and B.T. West (2015). ""Clustering by interviewer": a source of variance that is unaccounted for in single-stage health surveys." *American Journal of Epidemiology*, 182(2): 118-126.

Groves, R.M. and M.P. Couper (2012). *Nonresponse in household interview surveys*, John Wiley & Sons.

Kreuter, F. (2013). *Improving surveys with paradata: Analytic uses of process information*, Vol. 581, John Wiley & Sons.

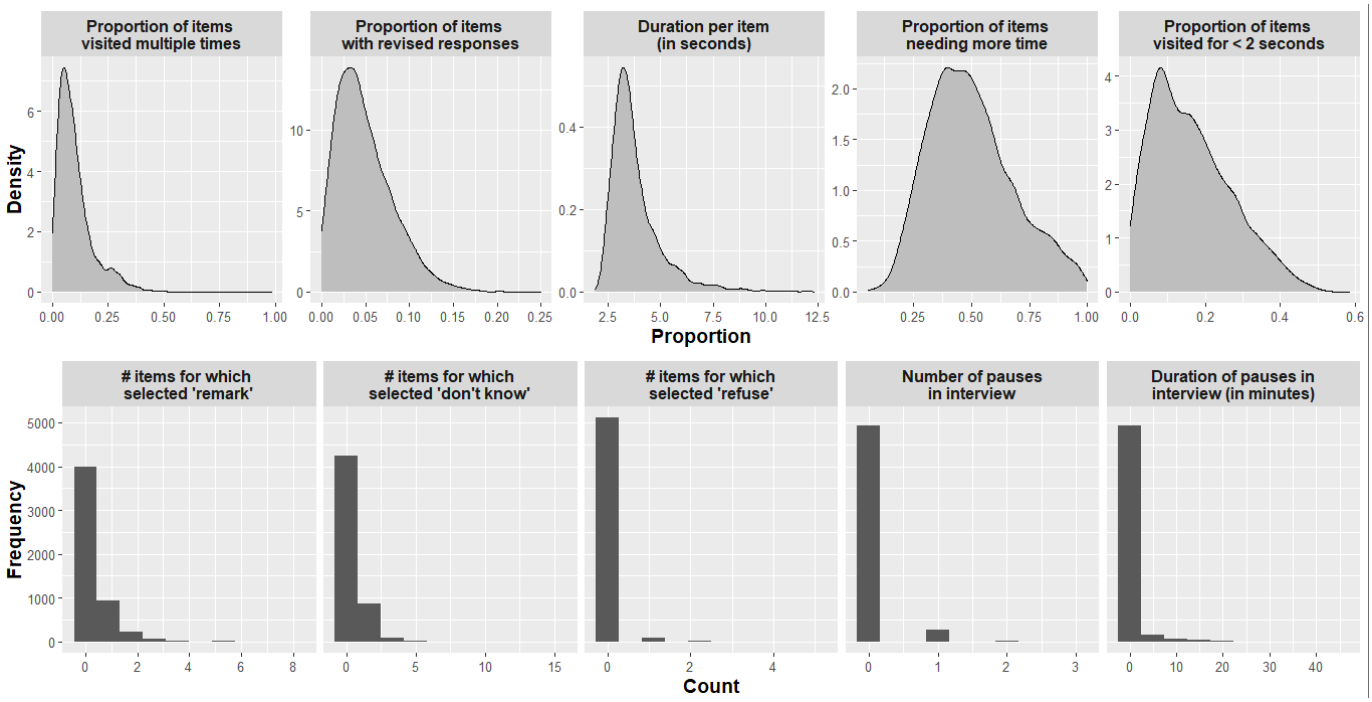
## Legend of Figures and Tables

**Figure 1. Distribution of item-wise summary of paradata for 200 common items across 8 sections**



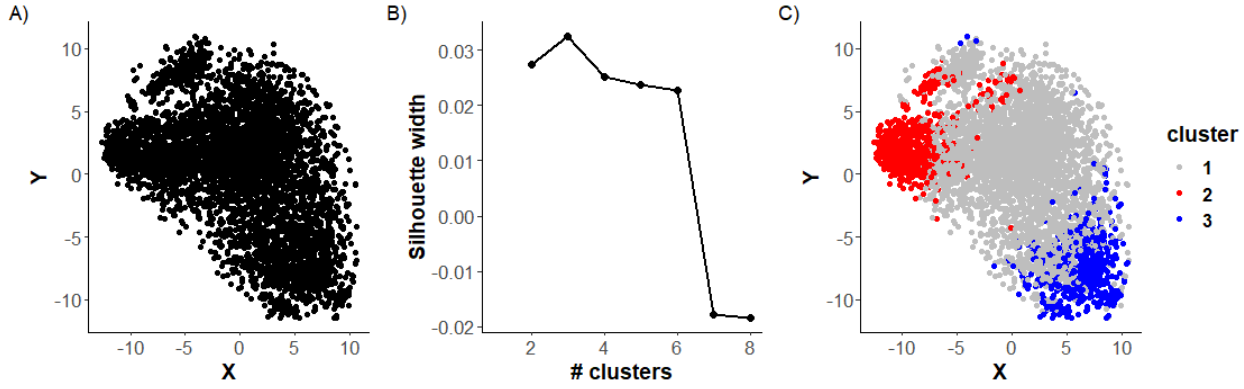

*The 200 items are plotted on x-axis in order that they appear in the questionnaire across the 8 sections. The y-axis presents the summary measure over 5231 interviews – median and interquartile range (IQR) of duration in seconds, total pause time in minutes, number of interviews in which the item was visited more than once or response was revised or maximum visit duration for item was less than 2 seconds, and the number of visits (denoted by #) for the remaining paradata indicators. The legend shows the 8 sections and the number in parentheses indicates the number of items out of the 200 common items in that section.*

**Figure 2. Distribution of interview-wise summary of paradata for 200 common items**



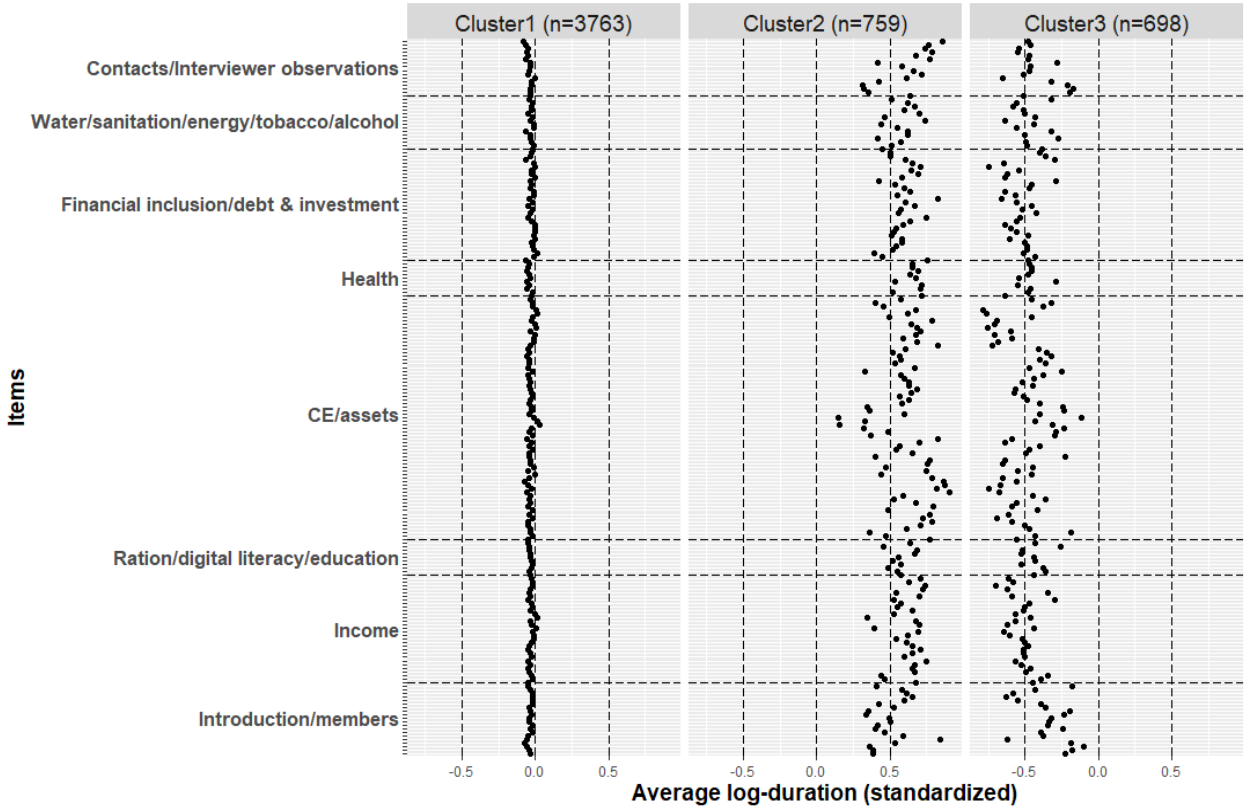

*For each interview we calculated a proportion or a count over 200 items and plotted its distribution across interviews. The top row presents the density plot for proportions across interviews and the bottom row presents the frequency distribution of the count data across interviews.*

**Figure 3. Visualisation of clusters and determining the optimal number of clusters**



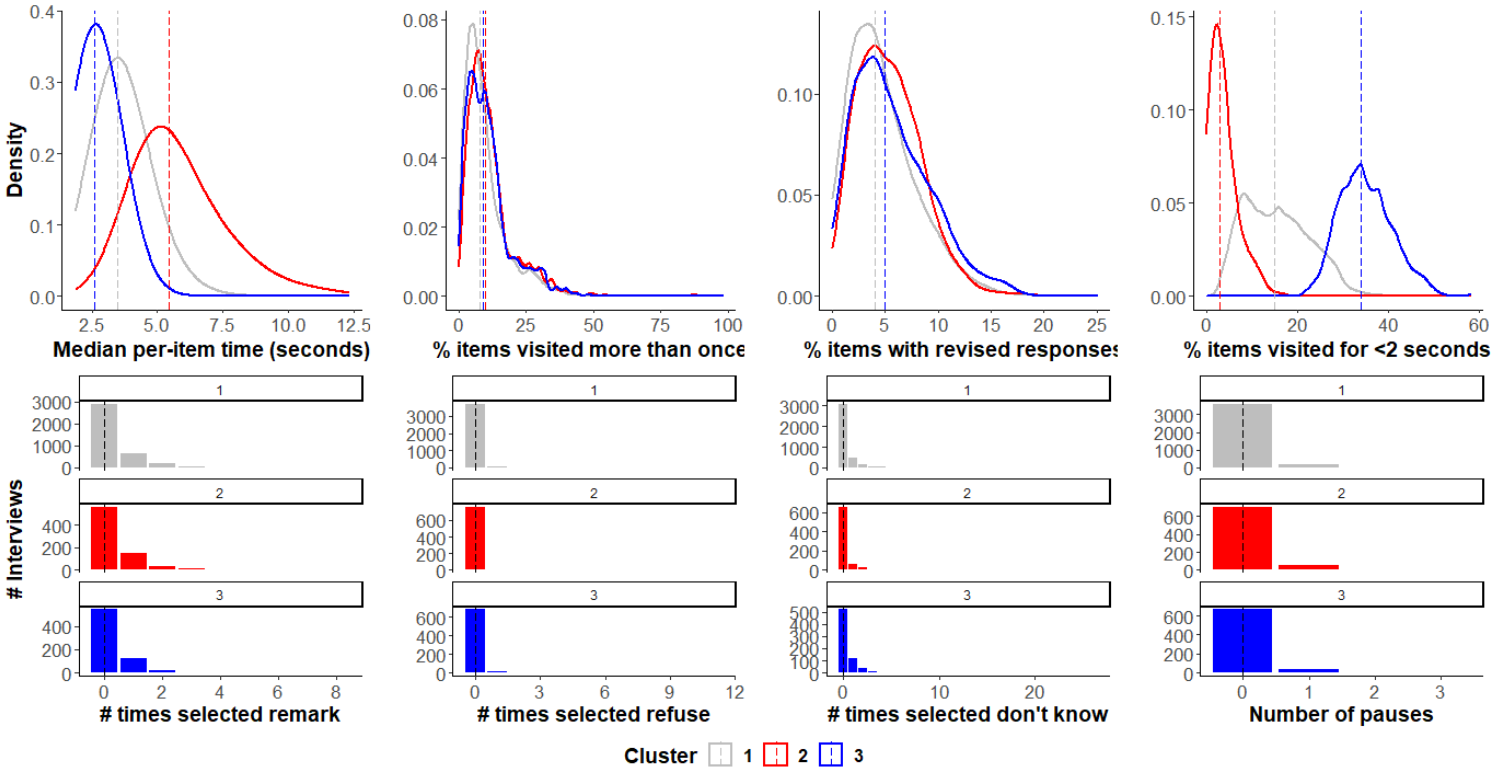

*A) 2-dimensional mapping of feature space using t-SNE. B) Silhouette method to determine the optimal number of clusters and C) Visualization of identified clusters in 2-dimensional space.*

**Figure 4. Average item-level duration for the 200 common items, by cluster**



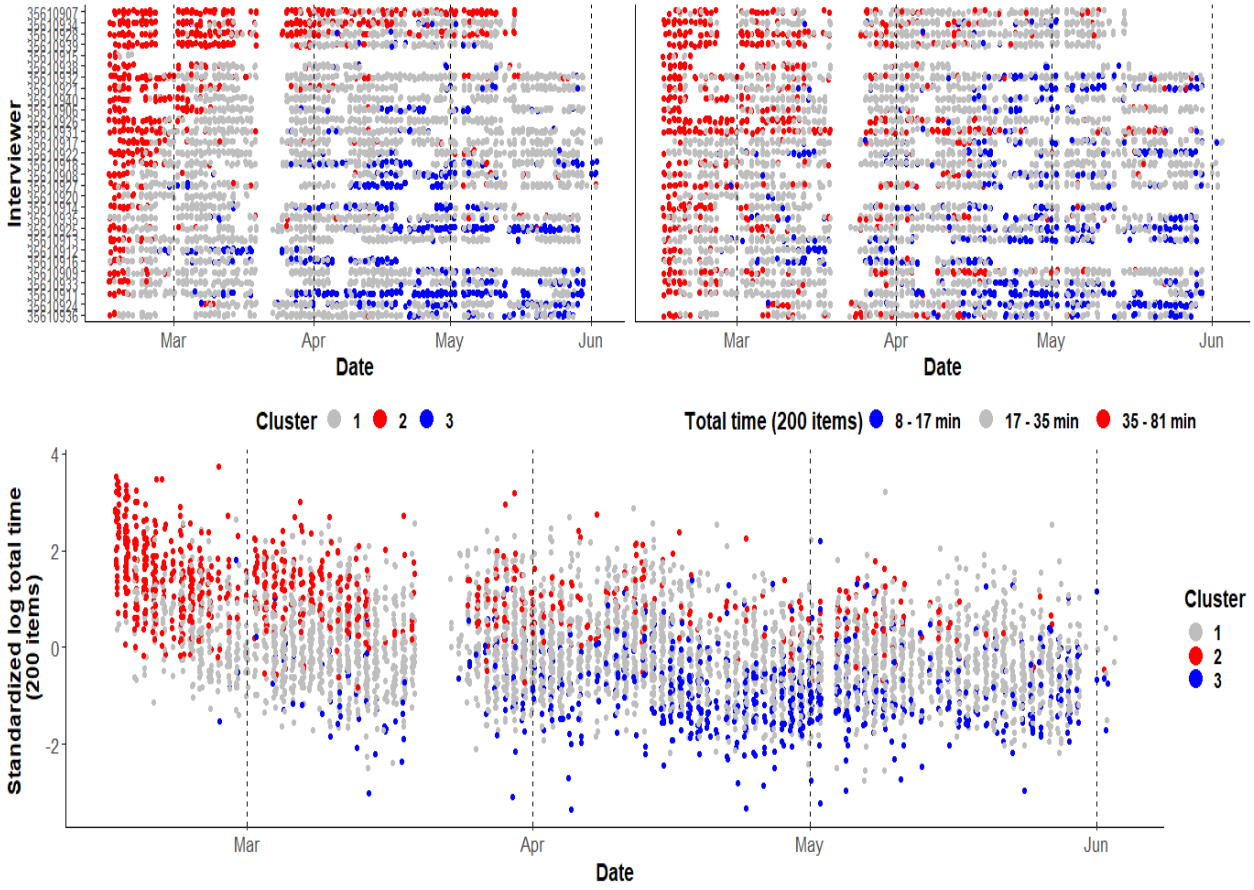

*Each point denotes the mean of standardised and log-transformed durations for an item, over all interviews in that cluster.*

**Figure 5. Distribution of interview-level summaries of paradata indicators across the 3 clusters**




*The interview-level summaries are based on the 200 common items across interviews.*

**Figure 6. Distribution of interviews across interviewers and over time, by cluster**




*Interviewers in top row plots are arranged in decreasing proportion of Cluster 2 interviews. Total time is total duration for 200 common items. In top right plot, it is categorized into three intervals – 698 interviews that took the least time, 8-17 minutes (denoted by blue), 3763 interviews that took 17-35 minutes (denoted by grey) and 759 interviews that took the most time, 35-81 minutes (denoted by red) to complete the 200 items.*

**Figure 7. Distribution of household characteristics of interviews, by cluster membership**

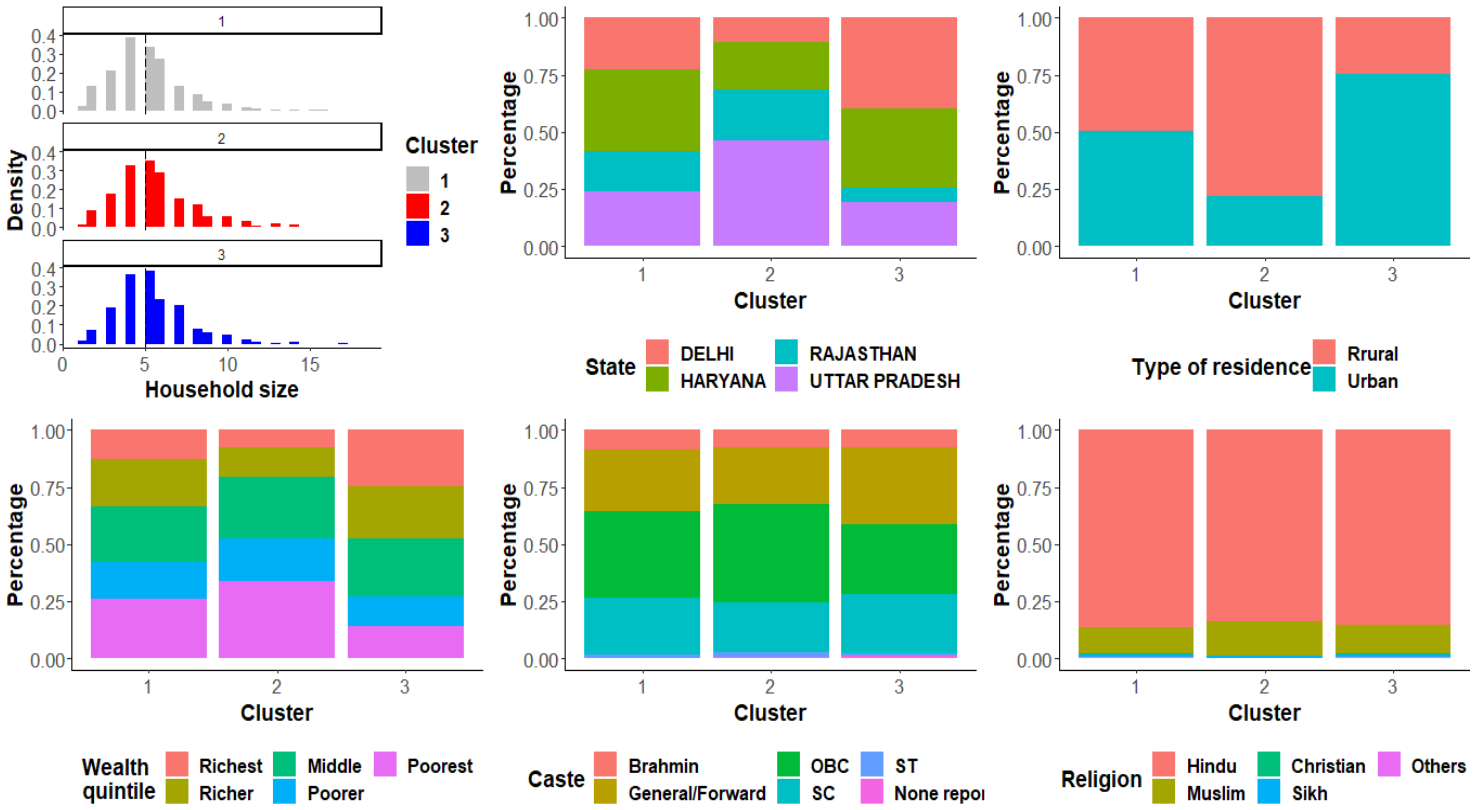


**Figure 8. Distribution of interviewer characteristics for the 3 clusters of interviews**

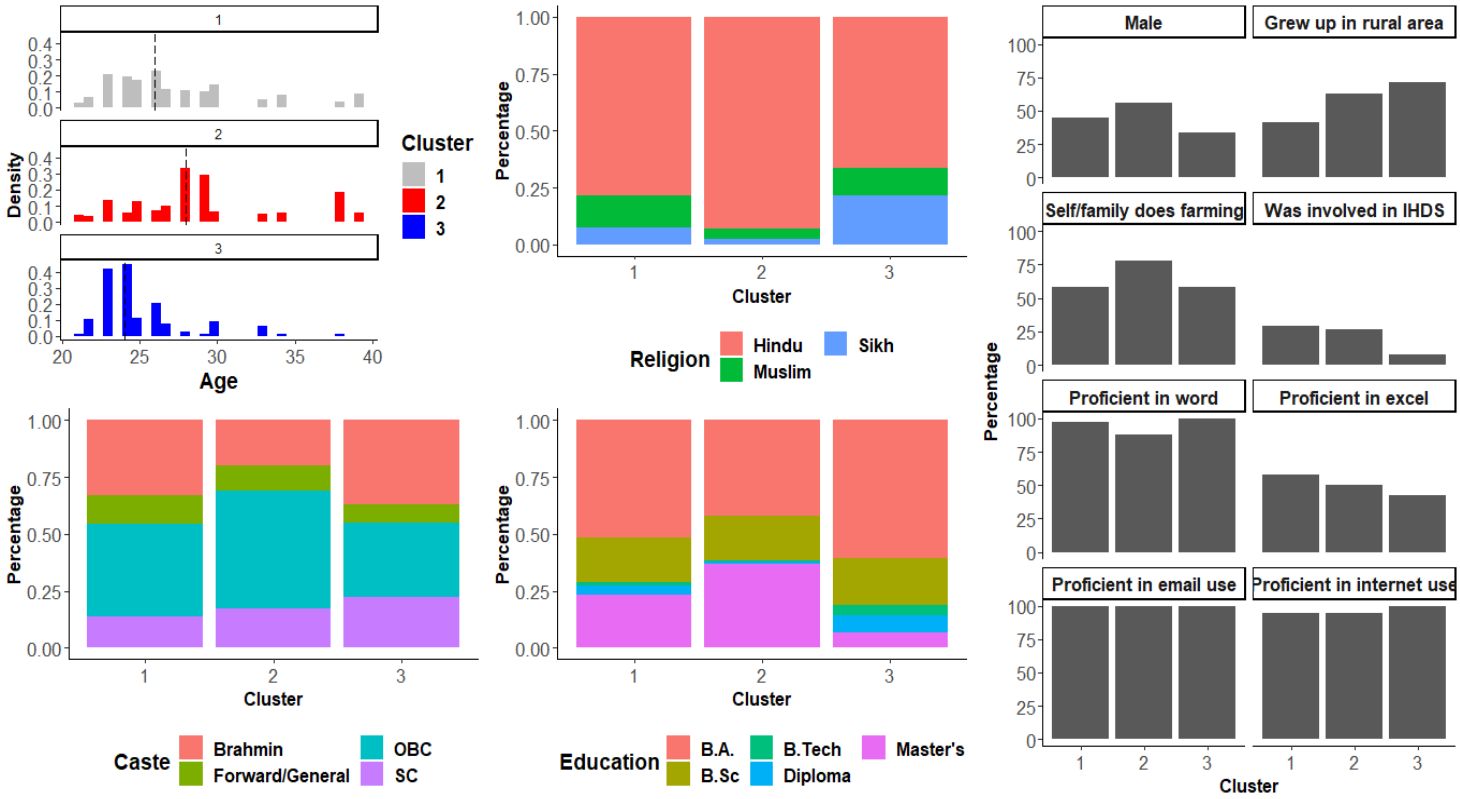
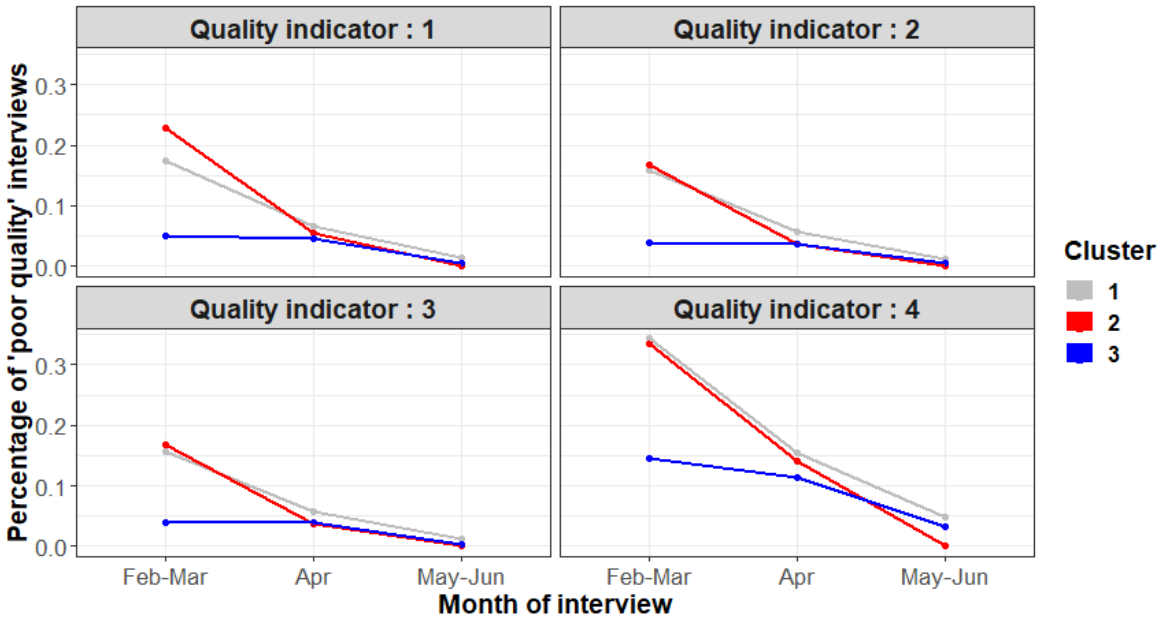


**Figure 9. Quality indicator and cluster membership**




*Interviews with mismatch for any household member between reported employment status in household member listing and detailed responses later in the interview regarding involvement in different activities, are tagged as 'poor quality' interviews.*

**Table 1. Characteristics of interviewers and households**

| **Interviewer characteristics** | **N (%)** | **Household characteristics** | **N (%)** |
|---|---|---|---|
| Total | 29 (100) | Total | 5220 (100) |
| Interviewer is female | 15 (51.7) | Household size, Median (Q1 – Q3) | 5 (4-6) |
| Age, Median (Q1 - Q3) | 26 (24 - 29) | State | |
| Education | | *Delhi* | 1216 (23.3) |
| *B.A./B.Sc./B.Tech./Diploma* | 22 (75.9) | *Haryana* | 173233.2 |
| *M.A./M.Com./M.S.W.* | 7 (24.1) | *Rajasthan* | 896 (17.2) |
| Religion | | *Uttar Pradesh* | 1376 (26.4) |
| *Hindu* | 24 (82.8) | Households in rural area | 2631 (50.4) |
| *Muslim* | 3 (10.3) | Household head caste | |
| *Sikh* | 2 (6.9) | *Brahmin* | 436 (8.4) |
| Caste | | *General/Forward* | 1435 (27.5) |
| *Brahmin* | 9 (31.0) | *Other backward class* | 1963 (37.6) |
| *Forward/General* | 4 (13.8) | *Scheduled caste* | 1287 (24.7) |
| *Other backward class* | 12 (41.4) | *Scheduled tribe* | 83 (1.6) |
| *Scheduled caste* | 4 (13.8) | *None reported* | 16 (0.3) |
| Interviewer grew up in rural area | 14 (48.3) | Household head religion | |
| Interviewer/family involved in agriculture | 18 (62.1) | *Hindu* | 4488 (86.0) |
| Interviewer has knowledge of | | *Muslim* | 622 (11.9) |
| Word | 28 (96.6) | *Others* | 110 (2.1) |
| Excel | 17 (58.6) | | |
| Email | 29 (100) | | |
| Internet | 28 (96.6) | | |
| Interviewer was involved in IHDS survey | 22 (75.9) | | |

**Table 2. Estimates of variance components from multilevel models of item-response times**

| Level | Null model | | Model 1, field characteristics | | Model 2, field, and household/interview characteristics | | Model 3, field, household/interview, and interviewer characteristics | |
|---|---|---|---|---|---|---|---|---|
| | Variance | ICC | Variance | ICC | Variance | ICC | Variance | ICC |
| Household/ interview | 0.0172 | 2.2 | 0.0184 | 2.9 | 0.0174 | 2.8 | 0.0174 | 2.8 |
| Village/town | 0.0044 | 0.6 | 0.0046 | 0.7 | 0.0044 | 0.7 | 0.0044 | 0.7 |
| Interviewer | 0.0183 | 2.3 | 0.0187 | 2.9 | 0.0189 | 3 | 0.0160 | 2.6 |
| Week into the survey | 0.0331 | 4.2 | 0.0344 | 5.4 | 0.0177 | 2.8 | 0.0177 | 2.8 |
| Residual (item) | 0.7238 | | 0.5668 | | 0.5668 | | 0.5668 | |

*Note*: Null model has random effects at levels: household/interview, village/town (i.e. primary sampling unit), week into the survey (23rd week was lumped with 22nd week) and interviewer.
Model 1: Null model + field characteristics – word count, type of field, sequence number in interview, and whether field has instructions.
Model 2: Model 1 + household/interview characteristics – state, type of residence, household size, household wealth quintile, religion & caste of household head, month of interview, and time of interview
Model 3: Model 2 + interviewer-level characteristics – sex, age, religion, caste, education, area in which interviewer grew up (rural/urban), whether interviewer or his/her family involved in agriculture, whether interviewer participated in IHDS survey, and knowledge of word, excel, email and internet

**Table 3. Effect estimates from final multilevel model including field, household/interview, and interviewer characteristics**

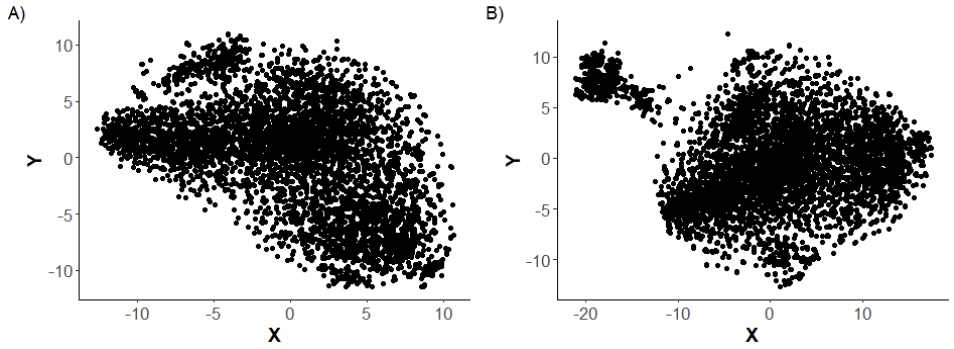
| Variable | Description | Summary | Estimate (SE) |
|---|---|---|---|
| **Field characteristics** | | **% of fields (n= 2572319 across all interviews)** | |
| Word count | Number of words in the field (including instructions) | Median (IQR): 19 (11-28) | 0.0017 (0.00001) *** |
| Sequence number | Counter indicating when item was asked in interview | Median (IQR): 261 (132-395) | -0.0003 (0.000003) *** |
| Field type | Single choice answer | 30.6% | |
| | Multiple choice answer | 2.7% | 0.6736 (0.003) *** |
| | Yes/No answer | 32.6% | -0.1618 (0.0012) *** |
| | Numeric | 24.6% | 0.3795 (0.0013) *** |
| | Real | 5.8% | 0.3719 (0.0021) *** |
| | Open | 0.2% | 0.9872 (0.0099) *** |
| | String | 3.5% | 0.9722 (0.0027) *** |
| Flag: interviewer instructions present | No | 40.9% | |
| | Yes | 59.1% | 0.1058 (0.0011) *** |
| **Household/interview characteristics** | | **% of households/ interviews (n=5231)** | |
| State | Delhi | 23.3% | |
| | Haryana | 33.2% | 0.0736 (0.0416) |
| | Rajasthan | 17.2% | 0.0621 (0.0454) |
| | Uttar Pradesh | 26.3% | -0.0371 (0.0393) |
| Type of residence | Urban | 49.6% | |
| | Rural | 50.4% | 0.0304 (0.0266) |
| Household size | Members who live under the same roof and share the same kitchen | Median (IQR): 5 (4-7) | -0.0022 (0.0012) |

| | | | |
|---|---|---|---|
| Household wealth quintile | Poorest | 25.6% | |
| | Poorer | 16% | 0.0174 (0.0064) ** |
| | Middle | 25% | 0.0139 (0.0061) * |
| | Richer | 19.7% | 0.0198 (0.0068) ** |
| | Richest | 13.7% | 0.0108 (0.008) |
| Religion of household head | Hindu | 85.9% | |
| | Muslim | 11.9% | -0.0084 (0.0076) |
| | Others | 2.1% | -0.0101 (0.0144) |
| Caste of household head | Brahmin | 8.4% | |
| | General/Forward | 27.5% | 0.0021 (0.0081) |
| | Other Backward Class | 37.6% | 0.0006 (0.008) |
| | Scheduled Caste | 24.6% | -0.0116 (0.0084) |
| | Scheduled Tribe | 1.6% | -0.0158 (0.0191) |
| | None reported | 0.3% | 0.0542 (0.0361) |
| Household engaged in farming activities | No | | |
| | Yes | | 0.0202 (0.0066) ** |
| Total number of fields filled in interview | 298 - 456 | 25% | |
| | 457 - 513 | 25.1% | 0.0373 (0.0058) *** |
| | 514 - 590 | 24.9% | 0.0494 (0.0067) *** |
| | 591 – 1105 | 25% | 0.0786 (0.009) *** |
| Month of interview | Feb | 12% | |
| | Mar | 27.8% | -0.073 (0.0329) * |
| | Apr | 33.1% | -0.2856 (0.0738) ** |
| | May | 26.8% | -0.2515 (0.0738) ** |
| | Jun | 0.2% | -0.3452 (0.0849) *** |
| Hour of day | Before noon | 20% | |
| | 12-3 pm | 66.5% | 0.0101 (0.005) * |
| | After 3 pm | 13.5% | -0.0238 (0.007) *** |
| **Interviewer characteristics** | | **% of interviewers (n=29)** | |
| Sex | Female | 51.7% | |
| | Male | 48.3% | 0.103 (0.0713) |
| Age | Age in years | Median (IQR): 26 (24 - 29) | 0.0044 (0.0094) |

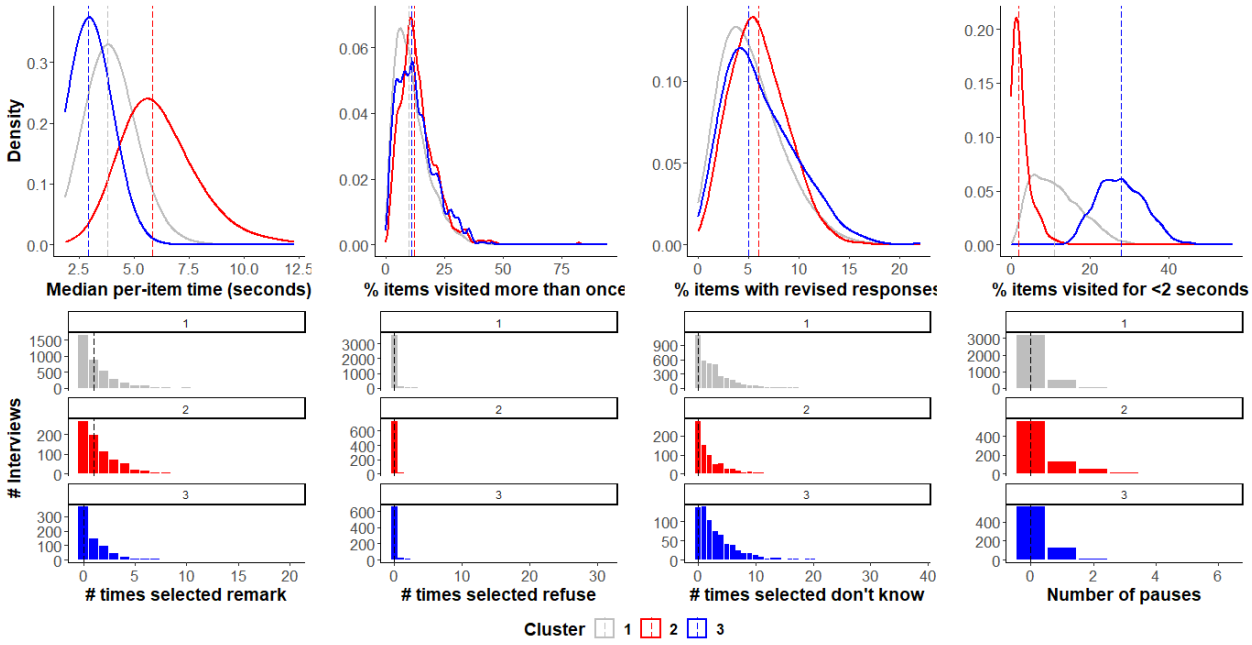
| | | | |
|---|---|---|---|
| Education | With Master's degree | 24.1% | |
| | Without Master's degree | 75.9% | -0.2109 (0.0975) * |
| Religion | Hindu | 82.8% | |
| | Muslim | 10.3% | 0.1793 (0.1126) |
| | Sikh | 6.9% | -0.1999 (0.1447) |
| Caste | Other Backward Class | 41.4% | |
| | Brahmin | 31% | 0.1795 (0.0819) * |
| | Scheduled Caste | 13.8% | 0.3523 (0.1365) * |
| | Forward/General | 13.8% | 0.2652 (0.1113) * |
| Type of area (rural/urban) in which interviewer grew up | Rural | 48.3% | |
| | Urban | 51.7% | -0.0553 (0.0734) |
| Whether interviewer/family involved in agriculture | No | 37.9% | |
| | Yes | 62.1% | 0.0908 (0.0737) |
| Knowledge of word, excel, and internet | Word – no vs. yes | 3.4% vs. 96.6% | -0.3337 (0.1627) |
| | Excel – no vs. yes | 41.4% vs. 58.6% | -0.0958 (0.0594) |
| | Internet – no vs. yes | 3.4% vs. 96.6% | -0.1372 (0.1688) |
| Whether interviewer was involved in IHDS survey | No | 75.9% | |
| | Yes | 24.1% | -0.071 (0.0819) |

# APPENDIX

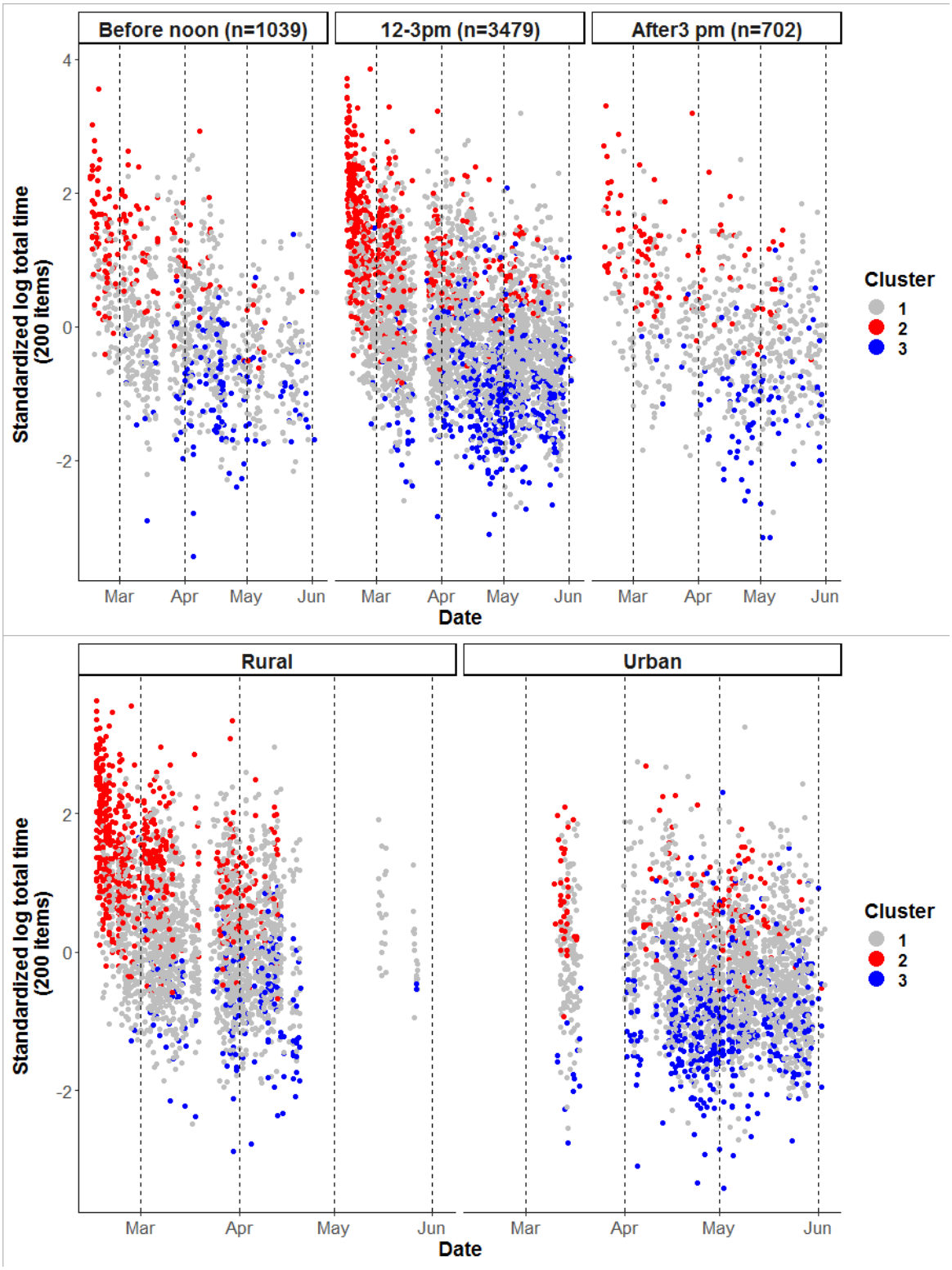
Two-dimensional mapping depends on distance matrix (choice of dissimilarity measure), which again depends on choice of features (type of paradata and how we define the feature/variable, that is, number of times item was visited during interview vs. whether item was visited multiple times). When we calculate Gower distance between two data points, it depends on the class of the variables – numerical or categorical. We present below t-SNE visualization of an alternative set of features created based on paradata for all 200 items – duration was kept as continuous, but all other paradata indicators were converted to a binary or variable. The plots demonstrate that choice and definition of features play an important role in identification of clusters.



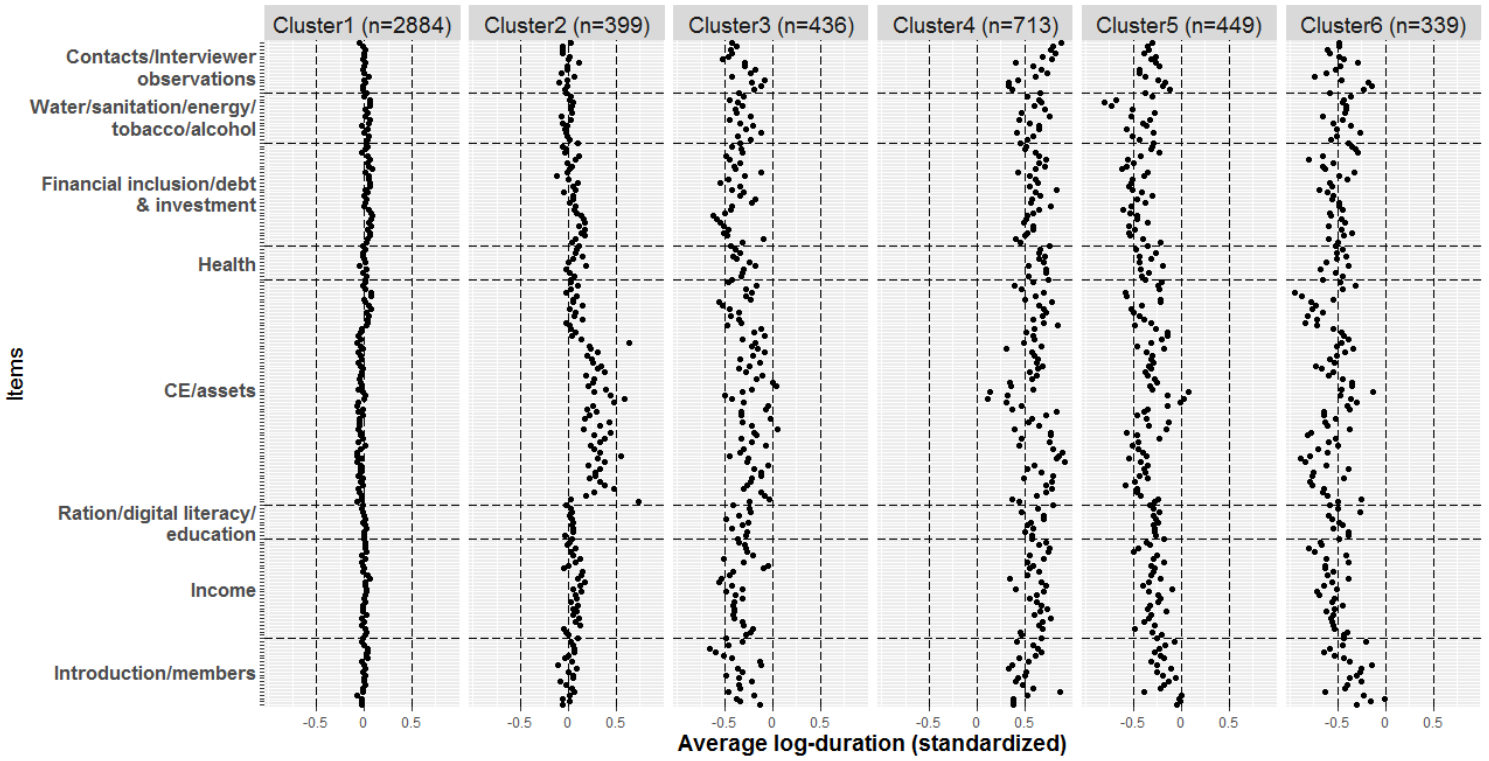

**Supplementary Figure 1.** t-SNE mapping of feature space, for two different sets of variables/features




**Supplementary Figure 2.** Distribution of interview-level summaries of paradata indicators across the 3 clusters. The interview-level summaries are based on all items in interviews.




**Supplementary Figure 3.** Distribution of interviews across the 3 clusters over survey period, by time of day and type of residence.




**Supplementary Figure 4.** Average item-level duration for 200 common items across 6 clusters. Each point denotes the mean of standardised and log-transformed durations for an item, over all interviews in that cluster.