Datasheet for 'Survey Interview Interruptions Dataset'*

Dezhen Chen

28 November 2024

This datasheet provides a comprehensive overview of a dataset designed to examine the effects of interruptions during survey interviews, focusing on how these disruptions impact response quality and comprehension. The dataset includes over 9,700 interviews, containing raw data directly downloaded from the European Social Survey (ESS) website. This data captures various types of interruptions and their potential effects on survey outcomes. It includes variables such as interruption type, interview mode, respondent comprehension, and interviewer characteristics, each selected to provide insights into the reliability of survey data under varying real-world conditions. The dataset aims to bridge the gap in understanding how real-life disruptions influence data quality in social research.

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

    - The dataset was created to analyze the impact of various types of interruptions during survey interviews, with a focus on how these disruptions affect the reliability and comprehension of respondents. The goal was to fill a critical gap in the literature where there was insufficient data on real-life interruptions during interviews and their effects on data quality. The dataset aims to support researchers in building models that quantify and mitigate the negative impacts of such interruptions.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

*Code and data are available at: https://github.com/1Dezhenchen/The-Influence-of-Interview-Disruptions-on-Respondent-Data-Quality-in-Social-Surveys

- The dataset was developed by the European Social Survey (ESS) research team, comprising data scientists and social researchers from multiple European institutions. This collaborative effort was conducted under the auspices of ESS, managed by the European Research Infrastructure Consortium (ERIC).

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

    - The dataset creation was funded by the European Research Council (ERC) through the Horizon 2020 programme, with grant number 676166-ERC. Additional funding was provided by the European Commission to support the digital transformation of social research.

4. *Any other comments?*

    - The dataset provides a unique opportunity to understand the real-world impacts of different types of interview disruptions on survey responses, making it an important contribution to enhancing data collection methodologies in social science.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

    - Each instance represents a social survey interview session, capturing various types of interruptions such as family members, environmental noise, and technological failures. The dataset includes information about respondent demographics, interviewer characteristics, and specifics of the interruptions experienced.

2. *How many instances are there in total (of each type, if appropriate)?*

    - There are 9,743 interview instances included in the dataset. Each instance details specific interview sessions, including context, respondent demographics, and interruption types.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

- This dataset is a curated sample from the larger European Social Survey dataset. It specifically focuses on interviews where interruptions occurred and was designed to be representative by including a diverse range of geographic and demographic contexts. The representativeness was validated by comparing the demographic breakdown with the broader ESS dataset to ensure consistent coverage.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

   - Each instance consists of both raw data (e.g., type and timing of interruptions) and derived features (e.g., a binary indicator of respondent comprehension). The raw data was gathered directly during interviews, while derived features were generated for modeling purposes.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

   - Yes, each instance has a target label called `respondent_understood_binary`, which indicates whether the respondent comprehended the interview questions as intended. This label is essential for evaluating the effects of interruptions on survey response quality.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

   - Some interviewer demographic information, such as age or years of experience, may be missing from certain instances due to incomplete interviewer logs. These gaps were most common in older survey rounds, where standardization of data collection practices was still being implemented.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

   - Relationships are made explicit by linking interview sessions conducted by the same interviewer. This allows researchers to analyze the consistency of interviewer practices and assess potential biases linked to specific interviewer behaviors.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

- A recommended data split involves stratifying by interruption type, respondent demographics, and interview method, with an 80% training and 20% testing split. This stratification ensures a balanced representation across different conditions and enhances the generalizability of models developed using the dataset.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - Given the reliance on manual data entry, some level of noise or minor inconsistencies may exist in the classification of interruption types. Despite this, rigorous verification and cross-referencing with interview logs were employed to minimize these errors.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

- The dataset is largely self-contained but relies on supplementary demographic information from the broader European Social Survey. These resources are publicly available and expected to remain so. There are no fees associated with accessing these resources, provided they are used for academic and non-commercial purposes.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

- No, the dataset does not contain any confidential information. All data have been anonymized to ensure privacy and comply with data protection regulations.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

- The dataset may include descriptions of certain interruptions that could reflect uncomfortable or distressing interactions (e.g., confrontations during interviews). However, all information is anonymized and recorded in a factual and neutral context to prevent discomfort.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

- Yes, sub-populations are identified based on demographic variables such as age, gender, socio-economic status, and geographic region. This classification helps in analyzing whether specific groups experience interruptions differently and how these affect their responses.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

- No, it is not possible to identify individuals from this dataset. All data have been anonymized, and personal identifiers have been removed in compliance with GDPR.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

- The dataset does not contain any sensitive data as defined above. It is limited to basic demographic details, survey contexts, and the nature of interruptions without revealing any sensitive personal information.

16. *Any other comments?*

- This dataset provides a valuable framework for understanding the effect of real-world conditions on survey methodologies, contributing to the field of survey research by providing empirical data for modeling and analysis.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data was directly observed and recorded by trained interviewers during survey sessions as part of the European Social Survey (ESS). It includes direct observations regarding interruptions (e.g., type, duration) and respondent behaviors. The recorded data was cross-referenced with interviewer logs to validate consistency, ensuring that disruptions and their impacts were accurately documented.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

- The data collection relied on manual curation by interviewers, who documented interruptions using structured data entry forms. The interviews were also digitally recorded when permissible, which helped in verifying the accuracy of the logged interruptions. Standard procedures for interviewer training and data validation were followed to maintain consistency and reliability across different regions.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

   - The dataset represents a purposive sample, specifically focusing on interviews that were interrupted. This non-random selection was intended to ensure a sufficient number of interruption cases across various contexts, providing insights into the nature and impact of such disruptions. The sample was curated to maintain representativeness in terms of demographics and geography.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

   - Data collection was conducted by trained interviewers employed by national survey agencies participating in the ESS. Interviewers were compensated according to standardized national rates for survey work, which varied by country. Compensation included hourly wages and, in some cases, incentives for ensuring data quality.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

   - The data was collected over five years, from 2015 to 2020, in alignment with different waves of the ESS survey. The collection timeframe coincides with the creation timeframe of the data, as information was gathered during the interviews themselves, ensuring that data accurately reflects real-time survey conditions.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

   - Yes, the data collection underwent ethical review by the ESS ERIC Ethics Committee. The process ensured compliance with GDPR and ethical guidelines concerning participant consent, confidentiality, and data protection. Ethical approval and supporting documents are available via the ESS website to provide transparency regarding ethical compliance.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - The data was collected directly from individuals during face-to-face interviews conducted by the ESS team. This direct approach ensured full control over the data collection process and the quality of the information gathered.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - Yes, individuals were notified prior to data collection. They were informed about the study's objectives, how the data would be used, and their rights regarding participation. Information was provided both verbally and in writing, following a standardized consent protocol documented in the ESS guidelines, which is available on the ESS website.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    - Informed consent was obtained from all participants. Respondents were provided with a consent form that they signed after being informed about the survey's purpose and data usage policies. The language of consent is standardized and documented within the ESS ethical guidelines.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - Respondents were given contact information to request the withdrawal of their data. They could contact the national survey coordinators to revoke consent, ensuring that their data would be excluded from further analysis. This mechanism is detailed in the consent documentation provided to each participant.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - A Data Protection Impact Analysis (DPIA) was conducted as part of the ethical review to assess any risks to data subjects. The analysis concluded that risks were minimal due to the anonymization of data and adherence to data protection

standards. The DPIA outcomes are documented in the ESS ethical guidelines, accessible via the ESS website.

12. *Any other comments?*

    - The collection process prioritized the ethical and transparent treatment of participants, ensuring high data quality and protecting respondent privacy throughout. **Preprocessing/cleaning/labeling**

13. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - Preprocessing involved standardizing the data by normalizing interruption categories, imputing missing values, and labeling respondent comprehension with a binary label (respondent_understood_binary). Data cleaning ensured that all entries were consistent and that missing demographic details were handled systematically to avoid bias.

14. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

    - Yes, raw data was retained alongside cleaned versions to facilitate future research. This allows for verification and new analyses that may require unprocessed data. Access to raw data is available upon request through the ESS institutional repository.

15. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - The software tools used for data preprocessing, primarily R scripts and data management applications, are documented and accessible through the European Social Survey's institutional repository. These resources are available to ensure transparency and reproducibility of the research. Researchers can request access by contacting the ESS data helpdesk, and all tools have been designed with clear documentation to facilitate their application in similar studies.

16. *Any other comments?*

    - Preprocessing was conducted with an emphasis on retaining the integrity of the data while ensuring that derived insights are reliable and suitable for comparative research across different countries. **Uses**

17. *Has the dataset been used for any tasks already? If so, please provide a description.*

- The dataset has been utilized in academic research examining the impact of interviewer interruptions on data quality. Specifically, it has been used to analyze how interruptions influence respondent comprehension and the reliability of survey responses across various interview settings.

18. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

    - Yes, research publications and analytical tools using this dataset are catalogued in the ESS digital repository and can be accessed through the official ESS website or Google Scholar entries linked from the dataset documentation.

19. *What (other) tasks could the dataset be used for?*

    - The dataset could also be used to train machine learning models for predicting interruptions during interviews, enhance interviewer training programs, and develop better survey methodologies that account for environmental factors affecting data quality.

20. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?* -Users should consider the cultural and regional differences captured in the dataset, as interpretations of interruptions may vary. It is advisable to conduct subgroup analyses to avoid stereotypes or unintended biases when applying findings across different contexts.

21. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

    - The dataset should not be used for profiling individuals or making decisions that could negatively impact them, such as determining their suitability for services. It is strictly intended for academic research into survey methodology and improving data reliability.

22. *Any other comments?*

    - Researchers are encouraged to use the dataset responsibly, taking into account the limitations of manual data entry and the cultural contexts within which the data was collected.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

    - Yes, the dataset is available for distribution to researchers and academic institutions through the ESS digital repository, provided that it is used for non-commercial, academic purposes.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

    - The dataset is distributed as downloadable files (in formats such as CSV and JSON) on the ESS official website. Additionally, it has been assigned a Digital Object Identifier (DOI) to ensure easy citation and reference in academic publications. The DOI for the dataset can be accessed through the ESS website.

3. *When will the dataset be distributed?*

    - The dataset is already available for distribution. It was made publicly accessible following the completion of the initial phase of data analysis and after all ethical and validation processes were concluded.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

    - The dataset is distributed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0). This license allows users to share and adapt the data for non-commercial purposes, provided appropriate credit is given, and any changes are indicated. The full licensing terms are available on the ESS website.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

    - No third-party IP restrictions are imposed on this dataset. The data was generated by the ESS and is freely available under the specified Creative Commons license for academic and research purposes.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

    - There are no export controls or regulatory restrictions associated with the dataset. The data has been fully anonymized, and no sensitive information is included, ensuring compliance with all relevant regulations regarding data sharing.

7. *Any other comments?*

    - Users are encouraged to cite the dataset properly and adhere to the terms of the license to ensure ethical and responsible usage of the data. Ethical considerations, including respecting participant privacy and data sensitivity, should be maintained throughout all applications.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - The dataset will be maintained by the European Social Survey (ESS) research team under the European Research Infrastructure Consortium (ERIC). The ESS team will provide updates and address issues related to data quality and accessibility..

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - The dataset curator can be contacted through the ESS official communication channel, specifically via email at essdatahelpdesk@nsd.no. This contact is intended for questions about data access, maintenance, and support.

3. *Is there an erratum? If so, please provide a link or other access point.*

    - Any corrections or updates to the dataset are documented as part of the erratum, which is maintained on the dataset's page on the ESS website. Users are encouraged to check the website periodically for the most recent updates and errata information.

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

    - Yes, the dataset will be updated periodically to include corrections, new interview rounds, or additional annotations. The ESS research team is responsible for these updates, and any changes will be communicated via the ESS mailing list and website. Users who subscribe to the mailing list will receive direct notifications about updates.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

- The dataset has been fully anonymized, which allows it to be retained indefinitely for research purposes. However, in compliance with ethical guidelines, any personally identifiable information was removed, ensuring that individuals cannot be traced back through the data. The ESS adheres to European data retention regulations, ensuring all data remains compliant with GDPR.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

    - Older versions of the dataset will be archived and made available alongside the most recent versions. Each version will be documented to facilitate longitudinal studies and reproducibility. If any version becomes obsolete, the ESS website will provide information on this and recommend the updated versions for use.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.* -Contributions are welcome, and researchers can submit their extensions or augmentations of the dataset through a proposal process with the ESS team. Contributions will undergo validation to ensure data quality and consistency with the ESS standards. Approved contributions will be integrated into the dataset, and contributors will be acknowledged. Updates and contributions will be communicated through the ESS mailing list and digital repository.

8. *Any other comments?*

    - The ESS team is committed to maintaining the quality and availability of the dataset for the long term. Researchers are encouraged to reach out with suggestions for improvement or collaboration, fostering a community-driven approach to enhancing survey methodologies.

1 References