gesis Leibniz Institute for the Social Sciences

Documentation of mail data collection

Sven Stadtmüller & Christoph Beuthner

June 2020, Version 1.0

Abstract

Transparency and reproducibility are key elements of good science, and this also holds for the process of data collection in scientific surveys. To conduct analyses based on survey data collected by others, researchers heavily depend on accurate documentation of all stages in the data collection process, either for generating new scientific evidence or for reviewing previous research findings (e.g., in replication studies). In this contribution, we propose documentation guidelines for mail surveys. In doing this, we not only focus on mail-only surveys but also cover documentation guidelines for self-administered mixed-mode surveys, thus taking into account their growing importance in the survey landscape.

Citation

Sven Stadtmüller and Beuthner, Christoph (2020). Documentation of mail data collection. Mannheim, GESIS - Leibniz Institute for the Social Sciences (GESIS- Survey Guidelines).

DOI: 10.15465/gesis-sg_en_032

Introduction

Transparency and reproducibility are key elements of good science, and this also holds for the process of data collection in scientific surveys. To conduct analyses based on survey data collected by others, researchers heavily depend on accurate documentation of all stages in the data collection process, either for generating new scientific evidence or for reviewing previous research findings (e.g., in replication studies).

In this contribution, we focus on documentation guidelines for mail surveys. As collecting data online has become increasingly popular, it appears that mail surveys come a bit out of fashion. While this is mainly true for surveys based on non-probability samples (e.g., opt-in panels) and for probability-based surveys of web-savvy populations with accessible email addresses (e.g., students), most researchers who aim to conduct probability-based population surveys still need to contact their target persons offline (e.g., via mail). This is because in many countries, probability samples from the general population are usually drawn from registration offices which only dispose of the names and postal addresses of their residential population but not of their email addresses. Furthermore, since possessing an email address is not mandatory, a part of the population can still not be reached via email. Additionally, contacting certain parts of the population (e.g. older persons) via email might not lead to the response rates desired (Dillman, Smyth, & Christian, 2014). However, to both exploit the cost-effectiveness of online surveys and the still existing preference in the general population to fill in a paper questionnaire instead of participating online (Mauz et al., 2018; Medway & Fulton, 2012), researchers increasingly rely on self-administered mixed-mode surveys (Biemer et al., 2018; de Leeuw, 2005). These also allow one to conduct probability sampling using the population register, and thus reduce under- or overcoverage (Gabler & Häder, 2016; Häder, 2016). In such mixed-mode surveys target persons can either complete an online or paper questionnaire. These surveys not only yield relatively low survey costs but also decent response rates (Greenlaw & Brown-Welty, 2009).

Apart from self-administered mixed-mode surveys growing in importance, single-mode mail surveys do also still exist in the survey landscape, albeit to a lesser extent than before the emergence of online surveys (ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V., n.d.). In sum, we expect mail surveys to be continuously part of the survey researcher's toolbox at least in the near future (Couper, 2011), thus underlining the importance of documentation guidelines for this survey mode. An extensive documentation allows researchers to reproduce research and to understand the research design in detail.

In the following, we propose mandatory information required for the documentation of mail surveys as well as other aspects covering additional information that may help researchers with specific interests. In doing so, we propose documentation guidelines relevant for mail-only surveys but also for self-administered mixed-mode surveys, thus taking into account their growing importance in the survey landscape. We arrange the relevant information for the documentation of mail surveys by splitting our recommendations into the parts "general information", "survey instrument", "recruiting", and "data processing".

General information

General information includes key aspects of the research project and the survey as a part thereof. In line with Schaurer, Kunz, & Heycke (2020), the following aspects are, in our point of view, mandatory ingredients of a good documentation:

☐ **Project Title**

The project title specifies the overarching aim of the research project the survey was conducted for. The project title may differ from the survey title since for convincing people to take part in the survey, it is often advisable to choose a short and appealing survey title.

☐ **Principal investigators, project team members and affiliations**

☐ **Relevant disciplines**

Here, it should be made clear for which disciplines the survey data may be mainly relevant (e.g., political science, public health, developmental psychology).

☐ **Funding**

If the researcher received external funding for the project and/or for the survey as a part thereof, documentation should include information about funding (e.g., by naming the funding organization). If not, it should be stated, that there was no (external) funding.

☐ **Implementation of the survey**

Documentation should include information on whether a fieldwork agency was commissioned for data collection (and if so, which one) or whether data collection was carried out by the researchers themselves (or by the project team). Moreover, it should be specified whether other third parties were involved in the process of data collection (e.g., print service providers).

☐ **Survey design**

It should be specified whether the survey was designed as a cross-sectional or longitudinal survey. In the case of a longitudinal survey, it should also be specified whether it is a trend or a panel survey and how many waves the longitudinal survey comprises.

☐ **Survey mode**

Concerning the survey mode, it should be made clear whether the mail was the only mode or whether additional modes (e.g., online) were used. In the case of self-administered mixed-mode surveys, it should be specified whether a simultaneous or a sequential mixed-mode design was implemented. While in the former, target persons can choose between filling in a paper questionnaire or taking part online from the very beginning, in sequential self-administered mixed-mode surveys the second mode (usually mail) is introduced at a later point in time (e.g., when contacting the target persons for the second time) (Tourangeau, 2017).

☐ **Target population and study area**

This includes information concerning the population the survey aims to make inferences about (e.g., the residential population of Germany or Hamburg aged 18 years and above).

This information is usually identical to the study/survey area. If not (e.g., if the target population are people who moved away from Hamburg between 2018 and 2019), the survey area should be specified as well.

☐ **Field time**

This information should include day, month, and year of the start and the end of the field time.

☐ **Sampling design**

Here, it should be documented whether a sampling design was implemented and, if so, whether a probability or non-probability sampling approach was chosen.

Moreover, the sampling design should be specified as detailed as possible. In the case of a non-probability sampling approach, this includes how respondents were recruited (e.g., via leaflets, advertisements, or via snowball methods).

In the case of a probability sampling approach, documentation should include information on (1) how the target population was defined (e.g., the residential population of Hamburg aged 18 years and above with German citizenship) (2) the source or provider of the sample (e.g., the registration office) and (3) whether a simple random sampling or a stratified random sampling and/or a cluster sampling approach was chosen.

In the case of cluster or multistage sampling (e.g., random route procedures), the sampling approach within each cluster/on the different stages should be specified as well. If, for instance, the target population consists of persons, but the sampling frame only covers households, it should be made clear how the selection process within the contacted households was organized (e.g., via the last-birthday-method). In the case of stratified random samples, the strata should be defined, and information should be provided on whether a proportionate or disproportionate random sampling approach was followed.

☐ **Sample size**

The net sample size should be reported in order to inform researchers about the absolute number of respondents.

☐ **Response metrics**

Response metrics provide key information about the success of the data collection process and should thus be provided as detailed as possible. To achieve this, careful documentation of the response and all response-related information from the field is crucial during the process of data collection.

The best way to document response metrics is to rely on the AAPOR Standard Definitions (American Association for Public Opinion Research, 2016). The proposed coding schemes by AAPOR were recently adapted to the German survey landscape (Stadtmüller et al., 2019). This contribution also includes tables with final disposition codes for mail surveys as well as a case study of a self-administered mixed-mode survey that illustrates how information from the field can be documented in this particular case.

When it comes to reporting response rates, it should be made clear how the response rate was calculated. Here again, the AAPOR Standard Definitions (2016) propose six different re-

sponse rates researchers can rely on. Accordingly, they should make explicit, which of these response rates they report.

☐ **Decision rules for response documentation**

When collecting survey data, researchers will encounter situations that require decision rules for response documentation that should be established in advance. Some examples for such situations requiring decision rules are (1) duplicate responses (e.g., when the researcher receives the first completed questionnaire only after having sent the reminder and receives the second completed questionnaire afterwards, or when a target person takes part in both survey modes) (2) changes in the state of eligibility after the start of the field time (e.g., when a target person dies or moves away after the start of the field time and the researcher becomes aware of this), and (3) responses after field time (i.e., how it was dealt with mail questionnaires that were returned after the defined end of field time). In our point of view, the best way to outline those decision rules is in a technical report. Additionally, this information can be included in a meta dataset to allow easy and fast access to relevant information concerning the data.

☐ **Data access**

Documentation should include information on whether survey data are accessible for others and, if so, where and how data can be obtained (e.g., in a data archive, on a website, or request). It is recommended that datasets are made accessible freely e.g. by publishing them in a data archive. Nevertheless, researchers should keep in mind that important reasons can stand against this, such as data protection regulations or moral reasons.

Apart from this mandatory information, the documentation of general aspects of the research project and the survey may also include the following:

☐ **DOI**

It is possible to assign a unique identification number (DOI) to the project documentation. This number allows other researchers to find the documentation more easily.

☐ **Sampling frame size**

If available, this information includes the number of elements in the sampling frame.

☐ **Sample characteristics**

Here, some information about the composition of the sample concerning key socio-demographics (e.g., gender, age, citizenship, educational level) may be provided. If available, these proportions can also be compared with the ones in the sampling frame, or even in the target population (e.g., by relying on census data).

☐ **Response metrics for subgroups**

Additional information concerning response metrics may include the reporting of response rates for socio-demographic subgroups (e.g., response rates by gender or age groups), or, in the case of self-administered mixed-mode surveys, the proportion of respondents for each survey mode. This information helps researchers who aim to carry out meta-analyses (e.g., to estimate the representation of socio-demographic groups in scientific surveys) since they can easily retrieve the required information without having to obtain the complete dataset.

☐ **Details related to response documentation**

Apart from the decision rules discussed above, there are other issues related to response documentation that may be of interest to researchers. This includes (1) whether an identifier was used and if so, how it was implemented (e.g., a string code on the title page of the survey) (2) how questionnaires, where identifiers were disguised or removed by respondents, were dealt with (3) how often returned questionnaires were delivered or picked up at the post office (e.g., on a daily basis) and (4) which additional response-related information was documented and integrated in the data set (e.g., date of response, mode of response (in self-administered mixed-mode surveys), or group indicators (e.g., experimental or incentive groups)).

Survey instrument

The heart of each survey is the questionnaire – and the best way for its documentation is to simply provide an electronic file showing the survey instrument in written form (e.g., a PDF file). In the case of various versions of the questionnaire (e.g., in cross-cultural surveys), a file for each version should be made public. In this file the questionnaire should be represented exactly in the form as it was used in the survey, including remarks with regard to the survey flow (filters). The following list contains key information about the questionnaire that should be provided as well.

☐ **Questionnaire topics**

This refers to a list of the main topics or modules of the questionnaire (e.g., political attitudes, demographics, mental health). These topics should best be ordered according to their sequence in the questionnaire.

☐ **The overall number of questions**

☐ **The overall number of items**

☐ **Sources and references**

Questionnaire documentation should include sources or references concerning items or scales used in the survey that were developed by other researchers.

☐ **Questionnaire versions**

As noted before, in the case of cross-cultural surveys, a file for each language version should be made public. Besides, different questionnaire versions may result from experiments implemented in the survey. In our point of view, it is not necessary to publish all versions unless they differ to a large extent. Rather, the number of questionnaire versions should be documented as well as the differences between the versions. Additional information may also include specifications regarding the experimental design (e.g., how many groups were implemented, procedures for the assignment of the target persons to the different groups, number of respondents in the different groups).

☐ **Codebook**

A codebook informing about the assignment of questions (to variables and variable labels) and answers (to value labels) should be stored within the documentation.

☐ **Meta- and Paradata**

It should be documented which kinds of meta- (e.g. experimental group, contact mode) and paradata (e.g. date of return, user agent) are stored in the dataset besides the questionnaire responses.

Besides, researchers may also provide the following information about the survey instrument:

☐ **Pretest**

This includes information on whether a pretest was conducted and, if so, which pretest technique was applied. Moreover, information about field time and sample size as well as the results and consequences drawn from the pretest concerning the final questionnaire may also be provided.

☐ **General design choices**

If general design choices were met beforehand, it is worthwhile to document them so that other researchers may obtain relevant information without having to work themselves through the questionnaire. To these design choices belong, for instance, the consistent usage of certain types of scales (e.g., uni- or bipolar, five-point or seven-point, fully labeled or partly labeled, horizontal or vertical arrangement) or decisions regarding filtering (e.g., input or output filters).

☐ **Implementation**

Information about survey implementation comprises, for instance, software solutions used for designing the questionnaire.

In the case of a self-administered mixed-mode survey, information about the survey instrument should be delivered separately for the paper and the online questionnaire. Documentation of online surveys is usually more extensive since it additionally requires information about, for instance, hosting or detailed information regarding technical implementations (e.g., whether a back-button or a progress indicator was used). For this, the Survey Guideline from Schaurer et al. (2020) provides useful and detailed recommendations.

We also encourage researchers conducting a mixed-mode survey to outline their strategy to deal with mode effects. Here, two main strategies can be distinguished: while the first one prioritizes minimizing mode effects by maximizing the similarity between the different instruments, the second strategy basically aims at minimizing the overall error by capitalizing on the strengths of each survey mode (Tourangeau, 2017).

Recruiting

Information about recruiting comprises all aspects of communication with the target population. An accurate documentation of all recruiting features is essential since they affect response metrics and, above all, the response rate of the survey. In our point of view, the following aspects are mandatory in the documentation of the recruiting strategy.

☐ **Number and timing of contacts**

This includes information on how often target persons were contacted. For each contact, documentation should also include information on (a) the timing of the contact (including

day, month, and year, and whether target persons were contacted in tranches) (b) rules for inclusion (i.e., who was contacted: all target persons or only those who had not yet responded?) (c) implementation of targeting methods (i.e., whether target persons were contacted in the same way, and if not, how contacts were targeted).

☐ **Contact letters**

(Anonymised) contact letters should be provided for each contact (as electronic files).

☐ **Additional material**

Researchers should outline whether additional material was sent to the target persons, and if so, which material was used (e.g., brochures, leaflets, data privacy sheet). We also encourage researchers to document additional material as electronic files.

☐ **Study title**

The study title communicated to the target persons should be specified.

☐ **Incentives**

Since they are known to heavily influence response rates, information about incentives (if they are used at all) should be as detailed as possible. This includes information about (a) the form of the incentive (i.e., prepaid vs. postpaid; monetary vs. non-monetary) (b) the value of the incentive and (c) whether incentives differed between contacts (e.g., larger incentives in the final reminder), target persons (e.g., larger incentives for certain socio-demographic groups), or, in the case of a self-administered mixed-mode survey, between modes of response (e.g., larger incentives for persons who participated online).

☐ **Mode for delivery**

This refers to information on whether postal delivery or other modes for delivery (e.g., students distributing the letters) were used. In the case of postal delivery, the documentation should include information on whether standard delivery or other modes for delivery (registered or customer post) were chosen. Moreover, researchers may document whether a stamp was fixed on the envelope or automatic franking was used. This information should be specified for each contact.

☐ **Mode for returning the questionnaire**

Moreover, researchers should outline how respondents were asked to return the questionnaire. In the case of postal delivery: Were respondents provided with a stamped return envelope or did they have to pay the postage by themselves?

☐ **Interventions**

This includes information on whether there were changes in the recruitment strategy during fieldwork (e.g., additional reminders, or adaptations of the incentive scheme).

Additional information regarding recruiting may also include the following aspects:

☐ **Website**

Researchers may also specify whether a study website existed. If so, they may provide a URL and briefly outline its contents (e.g., by providing screenshots).

☐ **Support**

This refers to information on whether target persons were provided with contact information for asking study-related questions (e.g., via a hotline, via email).

☐ **Material for delivery**

In order to provide others with detailed information about the delivery, the documentation may also include a scan of the mailing envelope. If this is not possible, information about its size and layout (e.g., whether logos were printed on the envelopes, and if so, which one/s) may be provided.

☐ **Material for return delivery**

In line with the former point, a scan of the envelope aimed for returning the questionnaire may also be of interest to some researchers. Apart from information about printed logos, it may be also relevant which address for return was printed on the envelope since some respondents may be confused if the questionnaire is returned to a third party (e.g., a print service provider).

Data processing

This area of documentation covers information related to all aspects of data processing, such as data entry, coding, and data preparation. In our point of view, the list of mandatory information provided in the course of the documentation of data collection in mail surveys comprises the following aspects:

☐ **Data entry process**

Researchers should specify how the process of data entry was implemented (e.g., automatic or manual entry). In the case of automatic data entry, the technical means should be documented (e.g., used devices and software). When data entry was done manually, it should be made clear how this process was organized (i.e., how many people were involved, whether data entry was carried out in tandem, and which software was used for generating the input mask and for data entry). Furthermore, researchers should outline whether quality checks of data entry were carried out, and if so, how these quality checks looked like.

☐ **Data entry rules**

This includes rules for data entry usually defined beforehand. Such rules may deal with (1) ambiguous responses (e.g., when respondents were asked to fill in their year of birth with four digits but answer “79”, or when respondents do not check a box but put a cross between two boxes) (2) filter errors (e.g., when respondents answer questions they should not have answered) (3) other types of inconsistent or implausible answers and (4) data entry rules for different types of missing values.

☐ **Coding of open-ended questions**

Here, researchers should outline how answers to open-ended questions were coded. This requires detailed information on the number of coders, the coding scheme and coder training (if applicable). If measures for coding were calculated (e.g., Cohens Cappa), this (and the respective values) should also be specified.

☐ **Generated variables**

If variables were generated based on respondents' answers, researchers should document how these variables were created and explain their purpose. The best way to illustrate the generation process is by publishing the respective code.

☐ **Weighting variables**

If weighting variables (e.g., design or calibration weights) are integrated in the data set, researchers should inform about their purpose and calculation.

☐ **Imputation**

If missing values in the data set were imputed, the mechanisms for imputation should be made clear.

☐ **Data linkage**

If the survey data were linked with data from other sources (e.g., administrative data), the linkage procedures should be outlined in detail, including (1) how consent to data linkage was obtained (2) how the additional data were obtained (3) which additional information was delivered (e.g., privacy forms) and (4) how the process of data linkage was implemented (e.g., identifiers used for data linkage).

☐ **Anonymization**

If variables in the data set have been anonymized due to data protection issues (e.g., the study identifier or information on respondents' country of birth), researchers should outline which variables were affected and how anonymization was realized.

References

ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V. (n.d.). *Jahresbericht*. Retrieved from https://www.adm-ev.de/jahresberichte/

American Association for Public Opinion Research. (2016). *Standard definitions: Final dispositions of case codes and outcome rates for surveys.* Retrieved from https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf

Biemer, P. P., Murphy, J., Zimmer, S., Berry, C., Deng, G., & Lewis, K. (2018). Using bonus monetary incentives to encourage web response in mixed-mode household surveys. *Journal of Survey Statistics and Methodology*, *6*(2), 240–261. https://doi.org/10.1093/jssam/smx015

Couper, M. P. (2011). The future of modes of data collection. *Public Opinion Quarterly*, *75*(5), 889–908. https://doi.org/10.1093/poq/nfr046

de Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, *21*(2), 233–255.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method* (4th edition). Hoboken: Wiley.

Gabler, S., & Häder, S. (2016). Sampling in theory. *GESIS Survey Guidelines*. https://doi.org/10.15465/GESIS-SG_EN_009

Greenlaw, C., & Brown-Welty, S. (2009). A comparison of web-based and paper-based survey methods: Testing assumptions of survey mode and response cost. *Evaluation Review*, *33*(5), 464–480. https://doi.org/10.1177/0193841X09340214

Häder, S. (2016). Sampling in practice. *GESIS Survey Guidelines*. https://doi.org/10.15465/GESIS-SG_EN_014

Mauz, E., Lippe, E. von der, Allen, J., Schilling, R., Müters, S., Hoebel, J., … Lange, C. (2018). Mixing modes in a population-based interview survey: Comparison of a sequential and a concurrent mixed-mode design for public health research. *Archives of Public Health*, *76*(1). https://doi.org/10.1186/s13690-017-0237-1

Medway, R. L., & Fulton, J. (2012). When more gets you less: A meta-analysis of the effect of concurrent web options on mail survey response rates. *Public Opinion Quarterly*, *76*(4), 733–746. https://doi.org/10.1093/poq/nfs047

Schaurer, I., Kunz, T., & Heycke, T. (2020). Documentation of online surveys. *GESIS Survey Guidelines*. https://doi.org/10.15465/GESIS-SG_EN_031

Stadtmüller, S., Silber, H., Daikeler, J., Martin, S., Sand, M., Schmich, P., … Zabal, A. (2019). Adaptation of the AAPOR final disposition codes for the german survey context. *GESIS Survey Guidelines*. https://doi.org/10.15465/GESIS-SG_EN_026

Tourangeau, R. (2017). Mixing modes: Tradeoffs among coverage, nonresponse, and measurement error. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, … B. T. West (Eds.), *Total survey error in practice* (pp. 115–132). https://doi.org/10.1002/9781119041702.ch6